L'agente razionale e i suoi limiti

Rinforzo, decisione e il problema dell'allineamento

Introduzione — La terza impossibilità

Ogni saggio di questo pentattico ruota attorno a una forma di impossibilità strutturale, e questa struttura non è ornamentale: è la tesi stessa dell'opera, il principio unificatore che attraversa tre secoli di matematica e tre decenni di storia dell'intelligenza artificiale. Il primo saggio ha documentato come i sistemi formali sufficientemente espressivi non possano essere al contempo completi e consistenti — l'impossibilità di Gödel, dimostrata nel 1931 con una tecnica di autoreferenza che ha il carattere dell'inevitabile — e come nessun algoritmo generale possa decidere la verità aritmetica, nella formulazione di Turing del 1936 che trasforma il paradosso logico in un problema sulla natura della computazione. Ha poi mostrato come queste limitazioni matematiche avessero un correlato ingegneristico preciso: il collasso dell'intelligenza artificiale simbolica di fronte alla complessità aperta del mondo reale non era un fallimento contingente, superabile con più regole o più potenza di calcolo, ma la manifestazione applicativa dell'impossibilità gödeliana di catturare strutture aperte con sistemi chiusi.

Il secondo saggio ha percorso il territorio dell'apprendimento: come un sistema costruisce, a partire da dati finiti, una rappresentazione della

Silvano Salvador

distribuzione di probabilità che governa il dominio. Ha mostrato che la distribuzione vera dei dati è inaccessibile in senso assoluto — avvicinabile soltanto asintoticamente da modelli parametrici finiti — e che questa distanza residua, misurabile in termini di divergenza di Kullback-Leibler dal prior universale di Solomonoff, non si annulla mai in un sistema di dimensione finita. La seconda impossibilità si manifesta nelle allucinazioni dei modelli linguistici, nell'opacità delle rappresentazioni distribuite, nella non-certificabilità formale dei sistemi statistici: tutte conseguenze della medesima struttura, la struttura dell'approssimazione irriducibilmente imperfetta.

Il terzo saggio porta alla luce una forma di impossibilità più vicina alla vita pratica e, per certi versi, più inquietante delle precedenti: la difficoltà fondamentale di specificare formalmente i valori di un agente intelligente. Non si tratta di una limitazione tecnica contingente, di un difetto nell'ingegneria dell'apprendimento per rinforzo o nella costruzione delle funzioni di ricompensa. Si tratta di un problema strutturale che affonda le proprie radici nella teoria della computazione, nella teoria della misura, nella teoria della scelta sociale e, in ultima analisi, nella filosofia morale. Un agente sufficientemente capace che ottimizza una funzione di ricompensa formale troverà inevitabilmente modi per massimizzarla che divergono dall'obiettivo che quella funzione intendeva approssimare. Questo non è un paradosso tecnico superabile con una specifica più accurata: è una proprietà della relazione tra ottimizzazione formale e valori umani, analoga alla relazione tra sistemi formali e matematica scoperta da Gödel.

Per comprendere perché questa impossibilità sia reale e non soltanto apparente — per distinguerla dalla difficoltà pratica di costruire buoni sistemi di rinforzo, che è un problema tecnico genuino ma di natura diversa — è necessario sviluppare l'intera matematica dell'agire razionale.

Silvano Salvador

Questa matematica comprende i processi decisionali markoviani e il loro apparato di programmazione dinamica, la teoria dei policy gradient e la geometria dell'informazione che la sottende, la teoria dei giochi e i suoi equilibri come struttura dell'interazione strategica, e infine il problema dell'allineamento nelle sue formulazioni più precise: la legge di Goodhart, la convergenza strumentale, il teorema di impossibilità di Arrow. Questo percorso occupa i cinque movimenti del presente saggio.

Il filo che attraversa tutti i movimenti è riconoscibile con chiarezza, sebbene si arricchisca di variazioni a ogni stazione. L'intelligenza artificiale simbolica aveva una visione dell'agire come deduzione: l'agente applica regole esplicite che codificano i suoi obiettivi e le sue conoscenze, e l'azione ottimale si deduce logicamente da queste regole con la stessa necessità con cui una conclusione segue da premesse in un sistema assiomatico. Il fallimento di questa visione — documentato nel primo saggio attraverso l'analisi del problema del frame, della chiusura per squalifica e della fragilità epistemica dei sistemi esperti — ha aperto la strada all'intelligenza artificiale basata sull'ottimizzazione: l'agente apprende una politica di comportamento che massimizza una ricompensa cumulativa attraverso l'interazione con l'ambiente. Questa visione è enormemente più flessibile e potente di quella deduttiva, capace di affrontare domini la cui complessità rende impossibile qualsiasi codifica esplicita della conoscenza rilevante.

Ma proprio questa potenza introduce il problema dell'allineamento in una forma nuova e più acuta. Laddove un agente deduttivo fallisce in modo prevedibile — non sa rispondere a situazioni fuori dalla propria base di conoscenza, e il fallimento è riconoscibile come gap nella copertura delle regole — un agente ottimizzante fallisce in modo imprevedibile e potenzialmente catastrofico. Esso trova soluzioni che rispettano la lettera della funzione di ricompensa mentre ne tradiscono lo

Silvano Salvador

spirito, massimizzando metriche formali attraverso comportamenti che nessun progettista aveva anticipato. La prevedibilità del fallimento simbolico era un costo ingegneristico; l'imprevedibilità del fallimento ottimizzante è una proprietà strutturale che tocca le fondamenta del progetto di costruire sistemi intelligenti allineati con i valori umani.

Questa terza impossibilità si inscrive nella struttura del pentattico come il momento dell'agire: dopo aver mostrato i limiti di ciò che si può provare (Saggio I) e di ciò che si può rappresentare e generare (Saggio II), il terzo saggio mostra i limiti di ciò che si può specificare come obiettivo. Le tre impossibilità sono variazioni sul medesimo tema — un tema che attraversa tutta la storia intellettuale del XX secolo e che l'intelligenza artificiale porta alla sua manifestazione più concreta e urgente: la finitezza e la formalizzazione non possono catturare completamente le strutture infinite e aperte della matematica (Gödel), del linguaggio e della percezione (paradigma informazionale di Shannon-Solomonoff), dei valori umani (Arrow-Goodhart). Il principio generale che le unifica potrebbe essere enunciato così: ogni sistema formale finito applicato a una struttura che lo trascende per apertura e complessità produce risultati parzialmente corretti e vuoti irriducibili che non possono essere colmati dall'interno del sistema stesso.

Vale la pena soffermarsi su una questione di metodo che riguarda l'intero pentattico e che in questo saggio acquista particolare rilievo. La matematica che viene sviluppata nelle pagine seguenti non è un ornamento erudito sovrapposto a un'analisi che potrebbe procedere senza di essa: è il mezzo attraverso cui le strutture profonde dei problemi diventano visibili. Senza l'equazione di ottimalità di Bellman, il concetto di politica ottimale rimane intuitivamente comprensibile ma non operazionalizzabile; senza la teoria del policy gradient, l'idea di ottimizzare direttamente il comportamento di un agente è un'aspirazione

Silvano Salvador

vaga; senza il teorema di impossibilità di Arrow nella sua formulazione precisa, la tesi che i valori umani non possano essere catturati da una funzione scalare resta un'intuizione filosofica contestabile. La matematica, in questo senso, non fonda l'analisi ma la rende rigorosa — la costringe alla precisione e rende impossibile nascondere le assunzioni, come Edmund Husserl aveva capito parlando di "fondazione" nel senso di Fundierung piuttosto che di Begründung.

La formalizzazione dell'agire: processi decisionali markoviani e programmazione dinamica

I. Dall'apprendimento all'azione: una distinzione ontologica

Il secondo saggio di questo pentattico si è occupato prevalentemente dell'apprendimento: come un sistema costruisce, a partire da dati, una distribuzione di probabilità che approssima la struttura del dominio — come un modello linguistico apprende la distribuzione dei testi naturali, come una rete convoluzionale apprende la distribuzione delle immagini naturali, come un modello generativo latente apprende la geometria della varietà sottostante ai dati. Ma l'apprendimento, per quanto necessario, non è sufficiente a caratterizzare un agente intelligente. Un agente non si limita a modellare il mondo: interagisce con esso, compie azioni che modificano lo stato dell'ambiente, riceve segnali di retroazione e persegue obiettivi che si dispiegano nel tempo. Questa dimensione

Silvano Salvador

temporale e interattiva richiede una matematica diversa da quella dell'inferenza: richiede la teoria delle decisioni sequenziali.

La distinzione tra apprendimento e agire non è soltanto funzionale ma ontologica, e merita di essere chiarita con precisione perché le sue implicazioni attraversano l'intero saggio. Un modello linguistico che prevede il prossimo token opera su dati che non può modificare: esso è, rispetto al proprio dominio di dati, un osservatore passivo che stima una distribuzione su una realtà che esiste indipendentemente dalla sua stima. Un agente che interagisce con un robot fisico, con un mercato finanziario, con un gioco da tavolo o con un utente umano è invece parte costituente dell'ambiente che modella: le sue azioni cambiano lo stato dell'ambiente, e lo stato futuro dipende dalla sequenza delle azioni passate. Questa circolarità — l'agente che influenza i dati su cui è addestrato, o più in generale i dati rilevanti per la valutazione delle proprie politiche — è la fonte delle difficoltà più profonde tanto sul piano tecnico quanto su quello filosofico del reinforcement learning e dell'intelligenza artificiale agentiva.

La circolarità ha una conseguenza formale immediata: le ipotesi statistiche che fondavano l'apprendimento nel secondo saggio — stazionarietà della distribuzione dei dati, indipendenza campionaria, separazione tra il sistema che apprende e la sorgente dei dati — vengono sistematicamente violate nell'agire. In un contesto di decisione sequenziale, la distribuzione degli stati futuri dipende dalla politica dell'agente, che a sua volta dipende dagli stati osservati in precedenza: il processo risultante è una catena di Markov controllata, non una sequenza di campioni indipendenti da una distribuzione fissa. La matematica appropriata è quella delle catene di Markov, della programmazione dinamica e della teoria dei punti fissi negli spazi di funzioni — strumenti che Richard Bellman sviluppò negli anni cinquanta in un contesto

Silvano Salvador

originariamente militare e di ricerca operativa, e che si rivelarono essere la struttura matematica fondamentale dell'agire razionale.

La separazione tra l'agente che apprende e l'ambiente con cui interagisce — separazione che negli MDP è catturata dalla distinzione tra politica π e funzione di transizione P — ha anche implicazioni per la filosofia della mente. Il paradigma cognitivista classico, erede del cartesianesimo, concepiva la cognizione come elaborazione interna di rappresentazioni simboliche da parte di un sistema chiuso: il cervello come calcolatore, il pensiero come manipolazione di strutture interne. Il paradigma dell'agente situated e embodied — elaborato da Rodney Brooks, Francisco Varela e Evan Thompson negli anni ottanta e novanta — sostituisce questa visione con una in cui la cognizione emerge dall'accoppiamento strutturale tra agente e ambiente: non c'è separazione netta tra "dentro" e "fuori", tra rappresentazione e azione. La matematica degli MDP cattura questa struttura in modo formalmente preciso: l'agente e l'ambiente sono entità distinte ma reciprocamente accoppiati attraverso il ciclo percezione-azione-ricompensa, e l'ottimizzazione si svolge su questo ciclo, non su rappresentazioni interne indipendenti dall'ambiente.

II. Il processo decisionale markoviano: struttura formale e proprietà fondamentali

Il quadro formale standard per la modellazione di decisioni sequenziali è il processo decisionale markoviano (Markov Decision Process, MDP), introdotto da Richard Bellman nel 1957 come generalizzazione della teoria delle catene di Markov alla presenza di un decisore attivo. Un MDP è una quintupla $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ i cui elementi meritano un'analisi

Silvano Salvador

separata tanto per la loro struttura matematica quanto per le assunzioni che ciascuno incorpora.

Lo spazio degli stati $\mathcal{S}$ è l'insieme di tutte le configurazioni possibili dell'ambiente, il repertorio completo di situazioni in cui l'agente può trovarsi. Può essere finito — come nel problema del labirinto o nel gioco del tris — numerabile infinito — come nello spazio delle sequenze di token in un modello linguistico trattato come ambiente — o continuo, come lo spazio delle configurazioni di un robot nello spazio fisico. La scelta di $\mathcal{S}$ incorpora già una decisione di modellazione profonda: cosa conta come lo "stato" dell'ambiente, quale informazione è rilevante per la decisione e quale può essere trascurata. In un gioco di scacchi, lo stato è la configurazione completa della scacchiera più il giocatore che deve muovere; in un sistema di controllo di un veicolo autonomo, lo stato include la posizione, la velocità, la mappa dell'ambiente e la storia recente delle osservazioni sensoriali.

Lo spazio delle azioni $\mathcal{A}$ è l'insieme di tutte le azioni che l'agente può compiere in ogni stato. Anch'esso può essere discreto o continuo, e la sua struttura determina in modo decisivo le proprietà computazionali del problema. Uno spazio di azioni discreto e finito — come l'insieme dei movimenti legali in una posizione scacchistica, o le quattro direzioni di movimento in un labirinto — permette il calcolo esplicito del massimo sull'azione in ogni aggiornamento. Uno spazio di azioni continuo — come i torque applicati alle giunture di un robot con n gradi di libertà, che vivono in $\mathbb{R}^n$ — rende questo calcolo impossibile per enumerazione e richiede metodi di ottimizzazione nell'azione stessa, una delle motivazioni principali per lo sviluppo dei metodi di policy gradient che costituiranno il tema della Parte Seconda.

Silvano Salvador

La funzione di transizione $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, dove $P(s' \mid s, a)$ è la probabilità che l'ambiente transiti nello stato s' quando l'agente compie l'azione a nello stato s, incorpora l'assunzione più importante della struttura MDP: la proprietà markoviana. Questa proprietà afferma che la distribuzione degli stati futuri dipende dallo stato presente e dall'azione corrente, ma non dalla storia degli stati precedenti:

$$P(s_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, \ldots, s_0, a_0) = P(s_{t+1} \mid s_t, a_t)$$

Lo stato s_t è una statistica sufficiente per predire la distribuzione degli stati futuri: tutta l'informazione rilevante sulla storia passata è condensata nello stato corrente. Questa è una restrizione forte che non vale per tutti i problemi reali — in molti contesti la storia passata è genuinamente rilevante oltre a quanto catturato dallo stato corrente, come in un problema di negoziazione dove il comportamento passato influenza la credibilità futura — ma che rende il problema matematicamente trattabile in modo che la formulazione non-markoviana non consente. Le estensioni ai processi decisionali parzialmente osservabili (POMDP), dove l'agente non osserva lo stato completo ma soltanto un'osservazione rumorosa di esso, e ai processi a memoria finita, dove si permette che la politica dipenda da una finestra temporale di stati passati, cercano di recuperare la realismo a costo di una maggiore complessità computazionale.

La funzione di ricompensa $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, dove $R(s, a)$ è la ricompensa scalare che l'agente riceve quando compie l'azione a nello stato s, è il punto di entrata dei valori nel formalismo degli MDP — e, come si argomentà nelle parti successive, il punto di maggiore vulnerabilità del framework dal punto di vista dell'allineamento. La ricompensa è scalare per necessità computazionale: un segnale vettoriale richiederebbe la

Silvano Salvador

specifica di un ordinamento o di una funzione di aggregazione tra le componenti, ripristinando il problema dell'aggregazione che la scalare aveva eluso. In alcune formulazioni più generali la ricompensa dipende anche dallo stato successivo: R(s, a, s'), catturando il fatto che la stessa azione può avere valori diversi a seconda di quale stato produce.

Il fattore di sconto $\gamma \in [0, 1)$ governa la preferenza dell'agente per le ricompense immediate rispetto a quelle future: una ricompensa ricevuta k passi nel futuro vale $\gamma^k \cdot R(s_t, a_t)$ in termini di valore al tempo t. Il requisito $\gamma < 1$ garantisce la convergenza della somma delle ricompense future anche su orizzonti temporali infiniti, e modella la preferenza per il presente che si osserva sia nei sistemi economici — il tasso di sconto intertemporale della teoria economica standard — sia nei sistemi biologici, dove la disponibilità di risorse certe nel presente è tipicamente preferita a disponibilità incerte nel futuro. Quando $\gamma = 1$, si parla di obiettivo a somma non scontata, il che richiede che il problema abbia un orizzonte finito T o che si utilizzi la ricompensa media come criterio di ottimalità.

Una politica π è una funzione che associa a ogni stato una distribuzione di probabilità sulle azioni: $\pi(a \mid s)$ è la probabilità che l'agente scelga l'azione a quando si trova nello stato s. Una politica deterministica, caso speciale del caso generale, seleziona un'unica azione per ogni stato: $a = \pi(s)$. L'obiettivo dell'agente è trovare la politica π^* che massimizza il ritorno atteso — la somma scontata delle ricompense future — a partire da ogni stato:

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s\right]$$

dove le aspettative sono prese rispetto alle transizioni stocastiche dell'ambiente (governate da P) e alla politica stocastica dell'agente (governata da π). Questa funzione $V^\pi : \mathcal{S} \to \mathbb{R}$ è la funzione di valore

Silvano Salvador

della politica π: essa assegna a ogni stato il valore — la ricompensa futura scontata attesa — di trovarsi in quello stato seguendo la politica π. La funzione di valore trasforma il problema dell'ottimizzazione sequenziale in un problema di calcolo su uno spazio di funzioni, e la programmazione dinamica di Bellman fornisce gli strumenti per risolverlo.

III. L'equazione di Bellman: un punto fisso nello spazio delle funzioni

Il contributo più importante di Richard Bellman alla teoria delle decisioni sequenziali non è l'introduzione degli MDP in quanto tale, ma la scoperta che la funzione di valore ottimale soddisfa un'equazione di ricorrenza — l'equazione di ottimalità di Bellman — che la caratterizza come punto fisso di un operatore contrattivo nello spazio delle funzioni. Questa caratterizzazione trasforma il problema dell'ottimizzazione sequenziale su orizzonti infiniti in un problema di iterazione di un operatore su uno spazio di Banach, con le garanzie di convergenza che il teorema del punto fisso di Banach garantisce per le contrazioni.

Per una politica arbitraria π, la funzione di valore soddisfa l'equazione di Bellman di consistenza, ottenuta espandendo la definizione di V^π per un passo temporale:

$$V^{\pi}(s) = \sum\{a \in \mathcal{A}\}\ \pi(a \mid s)\ [R(s, a) + \gamma \sum\{s' \in \mathcal{S}\}\ P(s' \mid s, a)\ V^{\pi}(s')]$$

Questa equazione esprime il valore di uno stato come la media, rispetto alle azioni secondo la politica e alle transizioni secondo il modello, della ricompensa immediata più il valore scontato dello stato successivo. La sua eleganza sta nel rendere manifesta la struttura ricorsiva della funzione di valore: il valore di uno stato dipende dai valori degli stati successori,

Silvano Salvador

che dipendono a loro volta dai valori dei loro successori, e così via. L'equazione di Bellman non è una semplice equazione da risolvere algebricamente: è una condizione di auto-consistenza che V^π deve soddisfare globalmente, per ogni stato simultaneamente.

In forma operatoriale, questa equazione può essere scritta come V^π = T^π V^π, dove T^π è l'operatore di Bellman associato alla politica π:

$$(T^{\pi} V)(s) = \sum\{a \in \mathcal{A}\} \pi(a \mid s) [R(s, a) + \gamma \sum\{s' \in \mathcal{S}\} P(s' \mid s, a) V(s')]$$

L'operatore T^π agisce sullo spazio delle funzioni limitate $\mathcal{S} \to \mathbb{R}$, che è uno spazio di Banach con la norma del sup $\|V\|_\infty = \max\{s \in \mathcal{S}\} |V(s)|$. La proprietà cruciale di T^π è che è una contrazione rispetto a questa norma, con coefficiente di contrazione γ:

$$\|T^{\pi} V - T^{\pi} U\|_\infty \leq \gamma \|V - U\|_\infty \text{ per ogni V, U}$$

La dimostrazione di questa disuguaglianza è diretta. Per ogni stato s:

$$|(T^{\pi} V)(s) - (T^{\pi} U)(s)| = |\gamma \sum_a \pi(a \mid s) \sum\{s'\} P(s' \mid s, a) (V(s') - U(s'))| \leq \gamma \sum_a \pi(a \mid s) \sum\{s'\} P(s' \mid s, a) |V(s') - U(s')| \leq \gamma \sum_a \pi(a \mid s) \sum\{s'\} P(s' \mid s, a) \|V - U\|_\infty = \gamma \|V - U\|_\infty$$

dove si sono usati la disuguaglianza triangolare, la non-negatività di π e P, e il fatto che $\sum_a \pi(a \mid s) \sum\{s'\} P(s' \mid s, a) = 1$. Prendendo il massimo su s si ottiene la contrazione di T^π.

Per il teorema di Banach del punto fisso, un'applicazione contrattiva su uno spazio di Banach completo ha un unico punto fisso, e le iterazioni convergono ad esso con velocità geometrica: $\|T^{\{\pi,n\}} V_0 - V^{\pi}\|_\infty \leq \gamma^n \|V_0 - V^{\pi}\|_\infty$. Questo significa che l'algoritmo di iterazione della politica — che parte da una funzione di valore iniziale arbitraria V_0 e applica ripetutamente T^π — converge a V^π per qualsiasi condizione iniziale, con velocità di convergenza determinata dal fattore di sconto γ.

Silvano Salvador

Più γ è vicino a 1, più lenta è la convergenza: quando l'agente valorizza molto il futuro, la propagazione del segnale di valore richiede molte iterazioni.

La funzione di valore ottimale V* è definita come il massimo su tutte le politiche:

$V^*(s) = \max_\pi V^\pi(s)$ per ogni $s \in \mathcal{S}$

e soddisfa l'equazione di ottimalità di Bellman, ottenuta sostituendo la media sulla politica con il massimo:

$$V^*(s) = \max_{a \in \mathcal{A}} [R(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V^*(s')]$$

L'operatore di ottimalità di Bellman T* è definito corrispondentemente:

$$(T^* V)(s) = \max_{a \in \mathcal{A}} [R(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V(s')]$$

e è anch'esso una contrazione con coefficiente γ. La dimostrazione è analoga a quella di T^π, con l'osservazione aggiuntiva che $|\max_a f(a) - \max_a g(a)| \le \max_a |f(a) - g(a)|$:

$$|(T^* V)(s) - (T^* U)(s)| = |\max_a [R(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s')] - \max_a [R(s,a) + \gamma \sum_{s'} P(s'|s,a) U(s')]| \le \max_a \gamma |\sum_{s'} P(s'|s,a)(V(s') - U(s'))| \le \max_a \gamma \sum_{s'} P(s'|s,a) \|V - U\|_\infty = \gamma \|V - U\|_\infty$$

Quindi V* è l'unico punto fisso di T*, e l'iterazione del valore $V_{k+1} = T^* V_k$ converge a V* per ogni condizione iniziale V_0, con garanzia:

$$\|V_k - V^*\|_\infty \le \gamma^k \|V_0 - V^*\|_\infty$$

Per raggiungere una precisione ε, sono sufficienti $k \ge \log(\|V_0 - V^*\|_\infty / \varepsilon) / \log(1/\gamma)$ iterazioni. Una volta nota V*, la politica ottimale deterministica si recupera come:

$$\pi^*(s) = \operatorname{argmax}_{a \in \mathcal{A}} [R(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V^*(s')]$$

Silvano Salvador

La struttura matematica qui esposta — l'operatore di Bellman come contrazione, la funzione di valore come punto fisso, la politica ottimale come argmax sull'equazione di Bellman — è il cuore della programmazione dinamica. Ma l'eleganza di queste equazioni non deve oscurare la difficoltà computazionale che nascondono. L'iterazione del valore richiede l'aggiornamento di $|\mathcal{S}|$ valori a ogni passo, ciascuno con un calcolo che coinvolge $|\mathcal{A}|$ azioni e, in linea di principio, l'intera distribuzione degli stati successori: la complessità per iterazione è $O(|\mathcal{S}|^2 \, |\mathcal{A}|)$ per MDP densi, intrattabile per spazi di stato di dimensione anche moderata. La vera sfida del reinforcement learning moderno è estendere queste garanzie teoriche ad ambienti dove $|\mathcal{S}|$ è astronomicamente grande o addirittura continuo — e questo richiede l'approssimazione funzionale e l'apprendimento dall'esperienza, temi delle sezioni successive.

IV. La funzione Q e il reinforcement learning model-free

Un'alternativa alla funzione di valore V(s) — che misura il valore di uno stato indipendentemente dall'azione intrapresa immediatamente dopo — è la funzione Q (o funzione valore-azione), che misura il valore di compiere una specifica azione in un dato stato, seguendo poi la politica π:

$Q^\pi(s, a) = R(s, a) + \gamma \sum_{s'} P(s' \mid s, a) V^\pi(s') = \mathbb{E}_\pi [\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s, a_0 = a]$

La relazione tra V^π e Q^π è semplice: $V^\pi(s) = \sum_a \pi(a \mid s) Q^\pi(s, a)$, ovvero il valore di uno stato è la media, secondo la politica, dei valori delle azioni disponibili in quello stato.

Silvano Salvador

La funzione di valore-azione ottimale Q* soddisfa l'equazione di ottimalità:

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s'} P(s' \mid s, a) \max_{a'} Q^*(s', a')$$

e la politica ottimale si recupera come $\pi^*(s) = \text{argmax}_a Q^*(s, a)$. La proprietà cruciale di Q* è che essa permette di agire in modo ottimale senza conoscere il modello di transizione P né la funzione di ricompensa R: è sufficiente conoscere Q* per selezionare l'azione ottimale in ogni stato. Questa proprietà rende Q* lo strumento fondamentale del model-free reinforcement learning — l'approccio che apprende direttamente dall'interazione con l'ambiente senza costruire un modello esplicito della sua dinamica.

L'equazione di ottimalità per Q* può essere scritta come il punto fisso dell'operatore T^*_Q:

$$(T^*_Q Q)(s, a) = R(s, a) + \gamma \sum_{s'} P(s' \mid s, a) \max_{a'} Q(s', a')$$

che è anch'esso una contrazione con coefficiente γ, con le stesse garanzie di convergenza. L'iterazione $Q_{k+1} = T_Q Q_k$ converge a Q per qualsiasi condizione iniziale Q_0.

V. Il Q-learning: apprendimento stocastico con garanzie di convergenza

Il Q-learning, introdotto da Christopher Watkins nella sua dissertazione di dottorato del 1989 e pubblicato formalmente con Paul Dayan nel 1992, è il primo algoritmo di reinforcement learning model-free con garanzie di convergenza a Q* sotto condizioni appropriate. L'idea è di approssimare l'iterazione di Q* usando aggiornamenti stocastici basati sulle singole transizioni (s, a, r, s') osservate durante l'interazione con

Silvano Salvador

l'ambiente, senza mai calcolare esplicitamente le aspettative sull'intera distribuzione di transizione.

La quantità centrale dell'aggiornamento è l'errore di differenza temporale (Temporal Difference error, o TD error):

$$\delta_t = r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)$$

Il TD error misura la discrepanza tra la stima corrente del valore $Q(s_t, a_t)$ e la stima di bootstrap $r_t + \gamma \max_{a'} Q(s_{t+1}, a')$ — la ricompensa immediata più il valore massimo stimato dello stato successivo. L'aggiornamento del Q-learning riduce questa discrepanza in modo proporzionale a un tasso di apprendimento $\alpha \in (0, 1]$:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \cdot \delta_t = Q(s_t, a_t) + \alpha [r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)]$$

Questo aggiornamento ha una struttura che merita di essere analizzata con cura. Il termine $r_t + \gamma \max_{a'} Q(s_{t+1}, a')$ è il target: la stima corrente del valore vero $Q^*(s_t, a_t)$, ottenuta usando la ricompensa osservata più il valore stimato del miglior successore. Il termine $-Q(s_t, a_t)$ è l'errore rispetto alla stima precedente. L'aggiornamento sposta la stima $Q(s_t, a_t)$ nella direzione del target con passo α.

La proprietà notevole del Q-learning è che il target usa il max sull'azione al passo successivo — non la media rispetto alla politica corrente, come farebbe l'aggiornamento SARSA. Questa caratteristica lo rende un algoritmo off-policy: converge alla Q-function ottimale Q^* indipendentemente dalla politica usata per raccogliere i dati, purché ogni coppia (s, a) sia visitata sufficientemente spesso. La convergenza di $Q(s, a) \rightarrow Q^*(s, a)$ per ogni (s, a) è garantita se i tassi di apprendimento $\alpha_t(s, a)$ soddisfano le condizioni di Robbins-Monro:

$$\sum_t \alpha_t(s, a) = \infty \text{ e } \sum_t \alpha_t^2(s, a) < \infty$$

Silvano Salvador

La prima condizione garantisce che l'algoritmo possa correggere gli errori iniziali; la seconda garantisce che gli aggiornamenti diventino sufficientemente piccoli da evitare oscillazioni. Condizioni tipiche che le soddisfano includono $\alpha_t = 1/t$ o $\alpha_t = 1/(t^{2/3})$.

La connessione neuroscientifica del Q-learning è uno dei risultati più belli nella storia delle scienze cognitive computazionali. Wolfram Schultz, Peter Dayan e P. Read Montague, in un articolo del 1997 che ha influenzato sia le neuroscienze sia il reinforcement learning, hanno mostrato che i neuroni dopaminergici nel mesencefalo dei primati emettono segnali di attivazione che corrispondono con straordinaria precisione matematica al TD error δ_t. In condizioni normali, questi neuroni sparano quando si riceve una ricompensa inattesa; dopo il condizionamento classico, sparano al momento dello stimolo predittivo della ricompensa — non al momento della ricompensa stessa — e mostrano un'inibizione quando la ricompensa attesa non arriva. Questo schema temporale di attivazione è esattamente quello predetto dall'aggiornamento TD: il segnale dopaminergico codifica la differenza tra la ricompensa attesa e quella ricevuta, propagandosi temporalmente all'indietro verso gli stimoli predittivi. Questa corrispondenza tra fisiologia e formalismo matematico è una delle evidenze più forti della tesi, sviluppata nel primo saggio, che l'intelligenza biologica operi attraverso meccanismi di ottimizzazione statistica piuttosto che logici.

VI. Le Deep Q-Networks: dalla tabella alla rete neurale

Nelle applicazioni di interesse pratico, lo spazio degli stati $\mathcal{S}$ è troppo grande per mantenere una tabella esplicita Q(s, a) per ogni coppia stato-azione. In problemi come il controllo di robot in spazi continui, la

Silvano Salvador

navigazione in ambienti ad alta dimensionalità, o il gioco di videogiochi da pixel grezzi, il numero di stati distinti è astronomico — dell'ordine di 10^{70} per il Go, e effettivamente infinito per gli spazi continui. La soluzione è approssimare la funzione Q con un approssimatore parametrico — nella versione contemporanea, una rete neurale profonda — e aggiornare i parametri della rete invece dei valori tabellari.

Le Deep Q-Networks (DQN) di Mnih e colleghi (2015), sviluppate a DeepMind, furono le prime a dimostrare che l'approccio scala a domini complessi senza ingegneria specifica per dominio. Addestrata soltanto dai pixel grezzi di giochi Atari e dal punteggio come segnale di ricompensa, la DQN raggiunse performance comparabili a quelle di un giocatore umano esperto su 49 giochi distinti — un risultato che nel paradigma simbolico avrebbe richiesto la progettazione di un sistema esperto separato per ciascun gioco, con conoscenza specifica del dominio accuratamente codificata da esperti umani.

L'architettura parametrica $Q(s, a; \theta)$ deve essere addestrata ottimizzando i parametri θ in modo che $Q(s, a; \theta)$ approssimi $Q^*(s, a)$. La funzione di perdita naturale è l'errore quadratico medio tra la stima corrente e il target di Bellman:

$$L(\theta) = \mathbb{E}_{(s, a, r, s') \sim \mathcal{D}} \left[(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta))^2 \right]$$

dove la specificazione del target con parametri θ^- distinti da quelli del valore corrente θ è una delle due innovazioni tecniche fondamentali di DQN, che ora si discutono nel dettaglio.

La prima innovazione è l'experience replay. Invece di aggiornare la rete con ogni transizione nell'ordine in cui viene osservata — il che introdurrebbe correlazioni temporali forti tra aggiornamenti consecutivi, destabilizzando il gradiente stocastico — le transizioni vengono memorizzate in un replay buffer $\mathcal{D}$ di dimensione fissa (tipicamente

Silvano Salvador

dell'ordine di 10^6 transizioni), e a ogni passo di aggiornamento si campiona uniformemente un mini-batch di transizioni. Il campionamento casuale spezza le correlazioni temporali e aumenta l'efficienza campionaria, permettendo a ogni transizione di contribuire a molteplici aggiornamenti indipendenti. Senza l'experience replay, l'addestramento di Q-network neurali su problemi sequenziali tende a instabilità e divergenza, perché le transizioni consecutive sono altamente correlate (stati vicini nel tempo sono spazialmente vicini nell'ambiente) e perché le distribuzioni degli aggiornamenti cambiano rapidamente man mano che la politica evolve.

La seconda innovazione è la target network: invece di usare la stessa rete θ sia per calcolare il valore corrente $Q(s, a; \theta)$ sia per calcolare il target $r + \gamma \max_{a'} Q(s', a'; \theta)$, si mantiene una copia separata θ^- dei parametri — la target network — aggiornata soltanto ogni C passi copiando i pesi della rete principale: $\theta^- \leftarrow \theta$ ogni C aggiornamenti. Tra un aggiornamento e il successivo, θ^- rimane fisso, il che stabilizza il target verso cui si sta ottimizzando. Senza questa separazione, ogni aggiornamento dei parametri θ modifica sia la stima del valore corrente sia il target verso cui ci si sta muovendo — un problema analogo alla rincorsa verso un obiettivo in movimento, che produce oscillazioni o divergenza.

Il gradiente della perdita rispetto a θ è:

$$\nabla \theta \, L(\theta) = -\mathbb{E}\{(s,a,r,s') \sim \mathcal{D}\} \left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta) \right) \cdot \nabla_\theta Q(s, a; \theta) \right]$$

dove il target $r + \gamma \max_{a'} Q(s', a'; \theta^-)$ è trattato come costante rispetto a θ durante la differenziazione — una scelta che semplifica il calcolo del gradiente e, in combinazione con la target network, stabilizza l'addestramento.

Silvano Salvador

VII. Il problema della sovrastima e le varianti di DQN

Le DQN e le loro varianti soffrono di un problema sistematico noto come sovrastima dei valori Q. Nell'aggiornamento del target, l'operatore $\max_{a'} Q(s', a'; \theta)$ usa la stessa rete sia per selezionare l'azione ottimale nello stato s' sia per stimarne il valore. Poiché le stime Q sono perturbate da rumore di stima (inevitabile data la natura campionaria dell'apprendimento), l'azione con il valore stimato massimo è frequentemente non quella con il valore vero massimo, ma quella che ha beneficiato di un errore di stima positivamente distorto. Applicare il max a valori stimati con rumore introduce una distorsione positiva sistematica: la stima del valore ottimale è in media superiore al valore vero, e questa sovrastima si propaga e si accumula attraverso le iterazioni dell'aggiornamento di Bellman.

Formalmente, se $Q(s', a'; \theta) = Q^*(s', a') + \varepsilon(s', a')$ con ε rumore di media zero e varianza σ^2 per ogni azione, allora:

$$\mathbb{E}[\max_{a'} Q(s', a'; \theta)] = \mathbb{E}[\max_{a'} (Q^*(s', a') + \varepsilon(s', a'))] \geq \max_{a'} Q^*(s', a') + \mathbb{E}[\max_{a'} \varepsilon(s', a')] \geq \max_{a'} Q^*(s', a')$$

con diseguaglianza stretta quando le varianze sono positive, perché la media del massimo di variabili casuali è maggiore o uguale al massimo delle medie (disuguaglianza di Jensen per il max).

Il Double Q-learning, introdotto da van Hasselt nel 2010 e adattato alle reti neurali da van Hasselt, Guez e Silver nel 2016, corregge questa distorsione usando due funzioni Q separate: la prima (con parametri θ) per selezionare l'azione ottimale, la seconda (con parametri θ^-) per stimarne il valore:

$$\text{target} = r + \gamma\, Q(s', \operatorname{argmax}_{a'} Q(s', a'; \theta); \theta^-)$$

Silvano Salvador

Poiché la selezione dell'azione avviene con θ e la valutazione con θ^-, i due errori di stima sono parzialmente incorrelati, e la distorsione sistematica si riduce drasticamente. In pratica si utilizzano la rete principale e la target network come le due funzioni Q, senza introdurre parametri aggiuntivi.

Un ulteriore raffinamento architetturale è la Dueling DQN (Wang et al., 2016), che decompone la funzione Q in due termini con significati distinti:

$$Q(s, a; \theta) = V(s; \theta_V) + A(s, a; \theta_A) - (1/|\mathcal{A}|) \sum_{a'} A(s, a'; \theta_A)$$

dove $V(s; \theta_V)$ è la funzione di valore dello stato — quanto è buono trovarsi nello stato s indipendentemente dall'azione — e $A(s, a; \theta_A)$ è la funzione di vantaggio — quanto l'azione a è preferibile rispetto alla media delle azioni disponibili in s. Il termine di sottrazione $(1/|\mathcal{A}|) \sum_{a'} A(s, a'; \theta_A)$ garantisce l'identificabilità della decomposizione: senza di esso, si potrebbe aggiungere una costante arbitraria a V e sottrarla da A senza modificare Q, rendendo il problema di ottimizzazione sottovincolato. Il modulo di identificabilità forza la media dei vantaggi a essere zero, rendendo V e A univocamente determinabili dai dati.

La motivazione per questa decomposizione è funzionale: in molti stati, la scelta dell'azione specifica è relativamente poco importante — ciò che conta è il valore dello stato in quanto tale. In questi contesti, la rete di vantaggio $A(s, a; \theta_A)$ apprende che le differenze tra le azioni sono trascurabili, mentre la rete di valore $V(s; \theta_V)$ apprende rapidamente il valore dello stato. Questo accelera la convergenza nelle situazioni — frequenti in ambienti reali — dove la scelta dell'azione ottimale è poco critica e il bottleneck dell'apprendimento è la stima del valore dello stato.

Silvano Salvador

VIII. Il reinforcement learning model-based: pianificazione, World Models e Dreamer

Il Q-learning e i suoi successori DQN sono metodi model-free: apprendono una politica o una funzione di valore direttamente dall'interazione con l'ambiente, senza costruire esplicitamente un modello delle sue dinamiche. Questa scelta è motivata dalla robustezza — non è necessario che il modello dell'ambiente sia accurato per imparare buone politiche — ma ha un costo in termini di efficienza campionaria: i metodi model-free richiedono tipicamente milioni di interazioni ambientali per convergere su problemi di complessità moderata. Il reinforcement learning model-based (MBRL) affronta questo trade-off dall'estremo opposto: investe nella costruzione di un modello esplicito dell'ambiente, poi usa quel modello per pianificare o per generare dati sintetici su cui addestrare la politica.

Il framework formale del MBRL aggiunge al MDP standard un modello appreso $\hat{T}(s' \mid s, a) \approx T(s' \mid s, a)$ e $\hat{R}(s, a) \approx R(s, a)$. La pianificazione con il modello appreso può avvenire con metodi di ricerca ad albero (come il Monte Carlo Tree Search di AlphaZero), con model-predictive control (MPC, che riottimizza la sequenza di azioni a ogni passo basandosi sulle predizioni del modello per un orizzonte finito H), o con Dyna-style methods che alternano aggiornamenti reali e simulati. Il guadagno in efficienza campionaria può essere di due o tre ordini di grandezza rispetto ai metodi model-free: nell'esperimento canonico di Dreamer su task di controllo continuo, Hafner e colleghi ottengono performance comparabili al SAC model-free con 100 volte meno interazioni reali.

World Models (Ha e Schmidhuber, 2018) costituisce il primo approccio sistematico a apprendere un modello del mondo latente di dimensione ridotta e a usarlo direttamente per l'addestramento di un controllore.

Silvano Salvador

L'architettura si articola in tre componenti. Il Visual Model V è un variational autoencoder che comprime le osservazioni visive x_t in rappresentazioni latenti $z_t = \text{enc}(x_t)$ di dimensione molto minore dell'input originale. Il Memory Model M è una Mixture Density Network accoppiata a una Recurrent Neural Network (MDN-RNN) che predice la distribuzione sullo stato latente futuro: $p(z_{t+1} \mid z_t, a_t, h_t)$, dove h_t è lo stato nascosto della RNN che comprime la storia. Il Controller C è una rete piccola che mappa lo stato (z_t, h_t) alle azioni $\pi(a_t \mid z_t, h_t)$. Il risultato chiave di Ha e Schmidhuber è che il controllore può essere addestrato interamente nel "sogno" — generando traiettorie nel modello M senza ulteriori interazioni reali — e poi trasferito nell'ambiente reale con performance competitive.

DreamerV3 (Hafner et al., 2023) porta questa linea di ricerca alla sua maturità attuale. Il modello latente usa un Recurrent State Space Model (RSSM) che fattorizza lo stato in una componente deterministica h_t (aggiornata dalla GRU: $h_t = f_\varphi(h_{t-1}, z_{t-1}, a_{t-1})$) e una componente stocastica z_t campionata da una categorica con probabilità $p_\varphi(z_t \mid h_t)$. L'addestramento minimizza la divergenza KL tra la distribuzione a posteriori $q_\varphi(z_t \mid h_t, x_t)$ e il prior $p_\varphi(z_t \mid h_t)$, penalizzata con il free-energy variazionale:

$$L_{\text{world}} = \mathbb{E}\left[\sum_t -\log p_\varphi(x_t \mid z_t, h_t) - \log p_\varphi(r_t \mid z_t, h_t) + \beta\, D_{KL}(q_\varphi(z_t \mid h_t, x_t) \,\|\, p_\varphi(z_t \mid h_t))\right]$$

che è l'ELBO applicato al modello latente sequenziale. La politica e la funzione di valore sono poi ottimizzate attraverso imagination rollouts di lunghezza H generati interamente dal modello latente: dato uno stato reale (h_τ, z_τ) come punto di partenza, si simulano H passi futuri usando il modello appreso, si calcola il ritorno λ-pesato lungo la traiettoria immaginata e si ottimizza con policy gradient. L'aspetto

Silvano Salvador

notevole di DreamerV3 è la sua universalità: gli stessi iperparametri senza adattamento specifico per dominio producono risultati state-of-the-art su 150 task eterogenei che spaziano dal controllo motorio continuo a giochi Atari a ambienti 3D, qualità che nessun metodo model-free aveva raggiunto.

Il trade-off fondamentale del MBRL è tra efficienza campionaria e compounding errors del modello. Ogni errore nella predizione $\hat{T}$ si propaga e si amplifica lungo l'orizzonte di pianificazione: dopo H passi di rollout, l'errore sulla distribuzione degli stati è proporzionale a $H \cdot \varepsilon_{model}$, dove ε_{model} è l'errore di predizione per passo. Questo rende i metodi MBRL instabili su orizzonti lunghi in ambienti con dinamiche complesse o alta stocasticità, mentre i metodi model-free che aggiornano su transizioni reali non soffrono di questo effetto di amplificazione. La soluzione pratica è l'uso di orizzonti di immaginazione brevi (H = 15-20 passi in DreamerV3) e di architetture latenti che comprimono l'errore mantenendo le caratteristiche rilevanti per il controllo.

Per il problema dell'allineamento, il MBRL introduce una proprietà strutturalmente rilevante: la possibilità di verifica offline della sicurezza. Un agente dotato di un modello del mondo esplicito può simulare il proprio comportamento futuro prima di agire, identificando sequenze di azioni che porterebbero a stati indesiderati e modificando la politica prima del deployment. Questa proprietà di look-ahead safety è assente nei metodi puramente model-free, dove le conseguenze di una politica si apprendono soltanto attraverso l'interazione effettiva. Il costo è la dipendenza dalla qualità del modello: un modello del mondo inaccurato produce valutazioni di sicurezza distorte, potenzialmente più pericolose dell'assenza di pianificazione se creano una falsa fiducia nella correttezza del comportamento pianificato.

Silvano Salvador

VIII-bis. L'offline reinforcement learning: CQL, IQL e Decision Transformer

Il reinforcement learning tradizionale è un metodo on-policy o off-policy che richiede interazione attiva con l'ambiente: l'agente esplora, commette errori e aggiorna la propria politica sulla base delle conseguenze reali delle sue azioni. In molti contesti ad alta posta — la medicina, l'aviazione, le decisioni finanziarie, il controllo di sistemi fisici — questa esplorazione interattiva è proibitiva per ragioni di sicurezza o costo. L'offline reinforcement learning (ORL) affronta questo vincolo apprendendo una politica ottimale da un dataset fisso D = {(s, a, r, s')} di transizioni raccolte in precedenza da una politica sconosciuta o da fonti eterogenee, senza ulteriori interazioni con l'ambiente.

La difficoltà fondamentale dell'offline RL è la distributional shift tra le azioni del dataset e le azioni valutate dalla politica appresa. In un aggiornamento standard di Q-learning, il target per l'aggiornamento di Q(s, a) include il termine max_{a'} Q(s', a'): questo massimo viene preso su tutte le azioni possibili in s', incluse quelle che non appaiono nel dataset. Per un agente che non può interagire con l'ambiente, queste azioni out-of-distribution possono ricevere valori Q artificialmente alti per via del bootstrapping — l'errore si propaga e si amplifica fino a rendere la politica appresa pericolosamente ottimistica rispetto alle transizioni mai osservate.

Conservative Q-Learning (CQL, Kumar et al., 2020) risolve questa difficoltà aggiungendo un termine di penalizzazione alla perdita di Q-learning standard che abbassa i valori Q per le azioni fuori distribuzione e li alza per le azioni nel dataset:

Silvano Salvador

$$L_{CQL}(Q) = L_{Bellman}(Q) + \alpha \cdot [E_{s \sim D}[\log \sum_{a'} \exp(Q(s, a'))] - E_{(s,a) \sim D}[Q(s, a)]]$$

Il primo termine del regolarizzatore riduce il soft-maximum dei valori Q su tutte le azioni, penalizzando le stime alte per le azioni non nel dataset; il secondo alza i valori Q per le coppie (s, a) effettivamente osservate. L'iperparametro α controlla la forza della conservatività. Kumar et al. dimostrano che CQL produce un lower bound rigoroso sul valore vero della politica: la politica appresa è pessimistica rispetto alle proprie capacità, il che la rende sicura da eseguire in deployment anche su transizioni mai osservate durante il training.

Implicit Q-Learning (IQL, Kostrikov et al., 2021) adotta un approccio diverso: invece di penalizzare le azioni out-of-distribution, evita completamente di valutarle. L'innovazione tecnica chiave è la separazione dell'aggiornamento della funzione di valore V(s) dall'aggiornamento della funzione Q(s, a). La funzione V(s) è addestrata come quantile regression sul dataset:

$$L_V(V) = E_{(s,a) \sim D}[L_\tau^{expectile}(Q(s,a) - V(s))], \quad \text{dove}$$
$$L_\tau^{exp}(u) = |\tau - 1(u < 0)|\, u^2$$

Con $\tau > 0.5$ questa perdita asymmetrica impara la V(s) come un quantile superiore dei valori Q nel dataset, senza mai richiedere la valutazione di azioni fuori distribuzione. L'estrazione della politica avviene poi con advantage-weighted regression: $\exp(\beta \cdot A(s,a)) \cdot \log \pi(a|s)$, dove $A(s,a) = Q(s,a) - V(s)$. IQL raggiunge performance competitive con i metodi on-policy su task di locomotion e manipulation partendo da dati eterogenei, e rappresenta attualmente l'approccio standard nell'offline RL per task continui.

Decision Transformer (Chen et al., 2021) offre una prospettiva radicalmente diversa: riformula l'offline RL come un problema di

Silvano Salvador

sequence modeling. Invece di imparare una funzione di valore, Decision Transformer addestra un transformer autoregressivo a predire l'azione a_t condizionata sulla storia di ritorni desiderati R̂_t, stati s_t e azioni passate. La condizione sul ritorno desiderato R̂_t permette di specificare la qualit' della politica a deployment time: impostando R̂_t al massimo ritorno osservato nel dataset si ottiene la migliore politica disponibile; impostando R̂_t a valori inferiori si ottengono politiche conservative.

La connessione con la prospettiva informazionale del Saggio II è diretta: Decision Transformer tratta l'offline RL come un problema di modellazione della distribuzione condizionale p(a_t | R̂_t, s_{≤t}, a_{<t}), esattamente come un modello linguistico modella p(token_t | context). Il dataset offline diventa il corpus di training; la politica ottimale emerge dal condizionamento sul ritorno desiderato, senza necessit' di ottimizzazione esplicita della funzione di valore. Questa equivalenza ha implicazioni per il problema dell'allineamento: un sistema di AI addestrato come sequence model può essere condizionato a produrre comportamenti allineati specificando il ritorno desiderato in termini di utilità umana, analogamente a come i modelli linguistici vengono condizionati con istruzioni.

X. Il limite del Q-learning e il paradosso della razionalità strumentale

Prima di procedere verso la teoria dei policy gradient, è necessario richiamare il contesto filosofico che rende il reinforcement learning non soltanto uno strumento ingegneristico ma un tema di riflessione fondamentale sul rapporto tra ottimizzazione e valori. Questo contesto non è un'aggiunta ornamentale all'esposizione tecnica: è il motivo per cui

Silvano Salvador

il presente saggio conclude, come i precedenti, con un'impossibilità strutturale piuttosto che con una soluzione tecnica.

Il paradigma simbolico dell'intelligenza artificiale aveva ereditato, dalla filosofia razionalistica della mente, una visione dell'agire come applicazione di regole: un agente razionale è un agente che conosce le regole del proprio dominio e le applica sistematicamente per raggiungere i propri obiettivi. In questa visione, la razionalità è essenzialmente procedurale — consiste nel seguire la procedura corretta — e i valori dell'agente sono codificati nelle regole stesse. Un sistema esperto per la diagnosi medica è "allineato" con i valori medici per costruzione: le sue regole codificano la conoscenza medica approvata dagli esperti, e l'agente non fa che applicarle. Il problema dell'allineamento, in questa visione, si riduce al problema di codificare correttamente le regole — un problema ingegneristico non banale, ma di principio risolvibile.

Il reinforcement learning rompe questa visione in modo radicale. L'agente non riceve regole: riceve una funzione di ricompensa scalare R(s, a), e il suo unico compito è massimizzare la somma scontata di questa ricompensa. La razionalità strumentale — la capacità di trovare mezzi efficaci per raggiungere un fine dato — viene separata in modo netto dalla specificazione del fine stesso. Un agente di reinforcement learning sufficientemente capace raggiungerà il fine specificato dalla funzione di ricompensa indipendentemente da come quel fine è stato specificato — e se la specificazione non cattura perfettamente l'obiettivo intenzionale del progettista, l'agente troverà modi per massimizzare la ricompensa formale che divergono dall'obiettivo reale.

Questa separazione — tra razionalità strumentale e specificazione degli obiettivi — è la fonte della terza impossibilità che questo saggio si propone di documentare. Non è un difetto tecnico dell'approccio ma una

Silvano Salvador

sua caratteristica strutturale: un sistema genuinamente ottimizzante è, per definizione, un sistema che trova ogni soluzione che massimizza la propria funzione obiettivo, incluse quelle che il progettista non aveva previsto. Quanto più capace è il sistema, tanto più completo è lo spazio delle soluzioni esplorate — e tanto più alta è la probabilità che alcune di esse divergano dall'intenzione originale, sfruttando imperfezioni della specifica che un ottimizzatore meno potente non riuscirebbe a trovare. Questo è il paradosso del progresso nel reinforcement learning: più capaci sono gli agenti, più grave diventa il problema dell'allineamento.

XI. Verso la teoria dei policy gradient: la geometria dell'ottimizzazione

L'approccio basato sulla funzione Q — le DQN e le loro varianti — è naturalmente adatto a spazi di azione discreti: il calcolo di max_{a'} Q(s', a') richiede di enumerare tutte le azioni possibili, operazione fattibile quando $|\mathcal{A}|$ è finito e moderatamente grande. Per spazi di azione continui — il controllo di un robot con articolazioni che variano in $\mathbb{R}^n$, la gestione di un portafoglio finanziario con allocazioni continue, la generazione token-per-token in un modello linguistico trattato come azione su uno spazio di vocabolario continuo — questo massimo non è calcolabile per enumerazione, e i metodi basati sulla Q-function richiedono estensioni significative come la DDPGarchitettura (Deep Deterministic Policy Gradient) di Lillicrap et al. (2016).

Esiste però un approccio alternativo che evita interamente questo problema ottimizzando direttamente i parametri della politica, senza intermediazione di una funzione di valore esplicita: i metodi di policy gradient. Questi metodi costruiscono un estimatore del gradiente dell'obiettivo di apprendimento direttamente rispetto ai parametri della

Silvano Salvador

politica, e aggiornano questi parametri seguendo il gradiente. Il prezzo è una varianza tipicamente più alta degli stimatori del gradiente rispetto ai metodi basati sul valore, e la necessità di tecniche sofisticate di riduzione della varianza — le baseline, le funzioni di critica, i metodi actor-critic — per rendere l'apprendimento praticamente convergente. Il beneficio è la generalità: i policy gradient si applicano a qualsiasi spazio di azione parametrizzabile, inclusi gli spazi continui ad altissima dimensionalità dei sistemi robotici e dei modelli linguistici contemporanei. Questi metodi costituiranno il tema della Parte Seconda.

La geometria dell'agire: policy gradient, ottimizzazione naturale e PPO

I. Ottimizzare la politica direttamente: la logica del policy gradient

L'approccio dei policy gradient abbandona l'intermediazione della funzione di valore — o, nei metodi actor-critic, la usa soltanto come strumento ausiliario per ridurre la varianza — e affronta il problema nella sua forma più diretta: trovare i parametri θ della politica π_θ(a | s) che massimizzano l'obiettivo di performance

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T} \gamma^t R(s_t, a_t)\right] = \mathbb{E}_{s_0 \sim d_0}\left[V^{\pi_\theta}(s_0)\right]$$

dove $\tau = (s_0, a_0, r_0, s_1, a_1, r_1, \ldots)$ è una traiettoria campionata seguendo la politica π_θ nell'ambiente con dinamica P, e d_0 è la distribuzione iniziale sugli stati. L'obiettivo $J(\theta)$ è la funzione di valore

Silvano Salvador

della politica π_θ mediata sulla distribuzione iniziale degli stati — il ritorno atteso dell'agente che segue la politica parametrizzata da θ.

La prima domanda è se $J(\theta)$ sia differenziabile rispetto a θ, e se sia possibile calcolare $\nabla_\theta J(\theta)$ in modo pratico. La risposta affermativa è non ovvia: $J(\theta)$ è un'aspettativa su traiettorie stocastiche che dipendono da θ in modo intricato — sia attraverso la politica $\pi_\theta(a \mid s)$ che distribuisce le azioni, sia (indirettamente, attraverso le azioni scelte) attraverso le transizioni dell'ambiente. Il teorema del policy gradient di Sutton, McAllester, Singh e Mansour (1999-2000) fornisce una formula sorprendentemente elegante per questo gradiente.

Il teorema del policy gradient afferma che, per qualsiasi politica differenziabile π_θ e funzione di valore V^{π_θ} associata:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s \sim d^{\pi_\theta},\, a \sim \pi_\theta(\cdot \mid s)} \left[\nabla_\theta \log \pi_\theta(a \mid s) \cdot Q^{\pi_\theta}(s, a) \right]$$

dove $d^{\pi_\theta}(s) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \pi_\theta)$ è la distribuzione degli stati visitati (con media geometrica scontata) seguendo la politica π_θ. Questa formula è il cuore di tutti i metodi di policy gradient, e la sua derivazione merita di essere seguita passo per passo per comprendere la struttura che essa rivela.

II. Derivazione del teorema del policy gradient

Si parte dalla definizione:

$$J(\theta) = \int_\tau p_\theta(\tau)\, R(\tau)\, d\tau$$

dove $p_\theta(\tau)$ è la probabilità della traiettoria τ sotto la politica π_θ e la dinamica P, e $R(\tau) = \sum_t \gamma^t R(s_t, a_t)$ è il ritorno totale della traiettoria.

Silvano Salvador

La probabilità di una traiettoria $\tau = (s_0, a_0, s_1, a_1, ..., s_T)$ si fattorizza come:

$$p_\theta(\tau) = d_0(s_0) \cdot \prod_{t=0}^{T-1} [\pi_\theta(a_t \mid s_t) \cdot P(s_{t+1} \mid s_t, a_t)]$$

Il gradiente di J è:

$$\nabla_\theta J(\theta) = \int_\tau \nabla_\theta p_\theta(\tau) R(\tau) d\tau$$

Per applicare il gradiente sotto il segno di integrale è necessario che l'integrazione e la differenziazione possano essere scambiate — una condizione tecnica soddisfatta sotto ipotesi di regolarità standard su π_θ. Si usa ora il trucco dell'identità del punteggio (score function trick o log-derivative trick):

$$\nabla_\theta p_\theta(\tau) = p_\theta(\tau) \cdot \nabla_\theta \log p_\theta(\tau)$$

che trasforma il gradiente della probabilità in un prodotto della probabilità per il gradiente del suo logaritmo. Sostituendo:

$$\nabla_\theta J(\theta) = \int_\tau p_\theta(\tau) \cdot \nabla_\theta \log p_\theta(\tau) \cdot R(\tau) d\tau = \mathbb{E}_{\tau \sim p_\theta} [\nabla_\theta \log p_\theta(\tau) \cdot R(\tau)]$$

Il logaritmo di $p_\theta(\tau)$ si decompone come:

$$\log p_\theta(\tau) = \log d_0(s_0) + \sum_{t=0}^{T-1} [\log \pi_\theta(a_t \mid s_t) + \log P(s_{t+1} \mid s_t, a_t)]$$

I termini $\log d_0(s_0)$ e $\log P(s_{t+1} \mid s_t, a_t)$ non dipendono da θ — la distribuzione iniziale e le transizioni dell'ambiente sono indipendenti dalla politica — e quindi il loro gradiente rispetto a θ è zero. Rimane:

$$\nabla_\theta \log p_\theta(\tau) = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t)$$

Sostituendo nell'espressione del gradiente:

Silvano Salvador

$\nabla \theta J(\theta) = \mathbb{E}\{\tau \sim p_\theta\} [\sum_\{t=0\}^\{T-1\} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot R(\tau)]$

Ora si applica la struttura causale del problema: l'azione al tempo t non può influenzare le ricompense ricevute prima del tempo t. Formalmente, $\nabla \theta \log \pi_\theta(a_t \mid s_t)$ *è incorrelato con le ricompense* $r\{t'\}$ per $t' < t$, dato s_t. Questo permette di sostituire $R(\tau) = \sum\{t'=0\}^\{T-1\} \gamma^\{t'\} r\{t'\}$ con soltanto le ricompense future a partire da t:

$\nabla \theta J(\theta) = \mathbb{E}\{\tau \sim p_\theta\} [\sum\{t=0\}^\{T-1\} \nabla \theta \log \pi_\theta(a_t \mid s_t) \cdot \sum\{t'=t\}^\{T-1\} \gamma^\{t'\} r\{t'\}]$

$= \mathbb{E}_\{\tau \sim p_\theta\} [\sum\{t=0\}^\{T-1\} \gamma^t \nabla \theta \log \pi_\theta(a_t \mid s_t) \cdot \sum\{t'=t\}^\{T-1\} \gamma^\{t'-t\} r\{t'\}]$

Il termine $\sum\{t'=t\}^\{T-1\} \gamma^\{t'-t\} r\{t'\}$ è esattamente il ritorno futuro G_t dalla posizione t. In aspettativa rispetto a traiettorie generate da π_θ, questo ritorno futuro coincide con $Q^\{\pi_\theta\}(s_t, a_t)$ — il valore della funzione Q alla coppia stato-azione (s_t, a_t). Si ottiene quindi la forma enunciata del teorema del policy gradient:

$\nabla \theta J(\theta) = \mathbb{E}\{s \sim d^\{\pi_\theta\}, a \sim \pi_\theta(\cdot|s)\} [\nabla_\theta \log \pi_\theta(a \mid s) \cdot Q^\{\pi_\theta\}(s, a)]$

Il termine $\nabla_\theta \log \pi_\theta(a \mid s)$ è la funzione score della politica — la direzione nello spazio dei parametri θ lungo la quale la probabilità di scegliere l'azione a nello stato s cresce più rapidamente. Il gradiente di J è la media pesata di queste direzioni, con pesi dati dal valore $Q^\{\pi_\theta\}(s, a)$: si spingono i parametri θ nella direzione che aumenta la probabilità delle azioni che hanno alto valore stimato, e si diminuisce la probabilità delle azioni a basso valore. Questa è la logica del reinforcement: rafforzare le azioni che portano a ricompense alte.

L'algoritmo REINFORCE (Williams, 1992) stima questo gradiente con campioni Monte Carlo. Si raccolgono N traiettorie complete $\tau^\{(1)\}$, ...,

Silvano Salvador

$\tau^{(N)}$ seguendo la politica corrente π_θ, si calcola il ritorno $G_t^{(i)}$ per ogni passo t di ogni traiettoria i, e si aggiornano i parametri:

$$\theta \leftarrow \theta + \alpha \cdot (1/N) \sum_{i=1}^{N} \sum_{t=0}^{T-1} \gamma^t \nabla_\theta \log \pi_\theta(a_t^{(i)} \mid s_t^{(i)}) \cdot G_t^{(i)}$$

Il principale difetto di REINFORCE nella sua forma grezza è l'elevata varianza dello stimatore del gradiente: i ritorni G_t variano enormemente tra traiettorie diverse, anche sotto la stessa politica, rendendo il gradiente stimato rumoroso e l'apprendimento lento e instabile. In problemi con ricompense sparse — dove la ricompensa è non nulla soltanto alla fine di lunghe traiettorie — la varianza diventa proibitiva e REINFORCE puro raramente converge in tempi ragionevoli.

III. Riduzione della varianza: baseline e funzione di critica

La varianza dello stimatore del gradiente può essere ridotta in modo sistematico sottraendo dal ritorno G_t una funzione baseline $b(s_t)$ — una funzione dello stato corrente che non dipende dall'azione scelta — senza introdurre distorsione nello stimatore. L'estimatore modificato diventa:

$$\nabla_\theta J(\theta) \approx (1/N) \sum_i \sum_t \gamma^t \nabla_\theta \log \pi_\theta(a_t^{(i)} \mid s_t^{(i)}) \cdot (G_t^{(i)} - b(s_t^{(i)}))$$

La prova che la sottrazione della baseline non distorce il gradiente è elegante e vale la pena riportarla esplicitamente. Per qualsiasi funzione $b(s_t)$ che non dipende dall'azione a_t:

Silvano Salvador

$\mathbb{E}_{a_t \sim \pi_\theta(\cdot|s_t)} [\nabla \theta \log \pi_\theta(a_t \mid s_t) \cdot b(s_t)] = \sum_{a_t} \pi_\theta(a_t \mid s_t) \cdot \nabla \theta \log \pi_\theta(a_t \mid s_t) \cdot b(s_t) = \sum_{a_t} \pi_\theta(a_t \mid s_t) \cdot (\nabla \theta \pi_\theta(a_t \mid s_t) / \pi_\theta(a_t \mid s_t)) \cdot b(s_t) = b(s_t) \cdot \sum_{a_t} \nabla_\theta \pi_\theta(a_t \mid s_t) = b(s_t) \cdot \nabla \theta \sum_{a_t} \pi_\theta(a_t \mid s_t) = b(s_t) \cdot \nabla_\theta 1 = 0$

dove si è usato il fatto che $\sum_{a_t} \pi_\theta(a_t \mid s_t) = 1$ è indipendente da θ. Quindi $E[\nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot b(s_t)] = 0$, e sottrarre $b(s_t)$ non cambia il valore atteso del gradiente ma può ridurne la varianza se b è scelto opportunamente.

La scelta ottimale della baseline — quella che minimizza la varianza dello stimatore del gradiente per ogni componente θ_i — è:

$b^*(s_t) = \mathbb{E}_{a_t \sim \pi_\theta} [\|\nabla \theta \log \pi_\theta(a_t \mid s_t)\|^2 \cdot Q^{\pi_\theta}(s_t, a_t)] / \mathbb{E}_{a_t \sim \pi_\theta} [\|\nabla_\theta \log \pi_\theta(a_t \mid s_t)\|^2]$

che è una media pesata del valore Q con pesi dati dalla magnitudine del gradiente della politica. Questa baseline ottimale è computazionalmente complessa da calcolare esattamente, ma in pratica si usa la funzione di valore $V^{\pi_\theta}(s_t)$ come approssimazione della baseline ottimale. Questa scelta porta naturalmente alla funzione di vantaggio:

$A^{\pi_\theta}(s_t, a_t) = Q^{\pi_\theta}(s_t, a_t) - V^{\pi_\theta}(s_t)$

che misura quanto l'azione a_t è migliore o peggiore del comportamento medio della politica π_θ nello stato s_t. Il gradiente stimato diventa:

$\nabla \theta J(\theta) \approx \mathbb{E}_{s, a \sim \pi_\theta} [\nabla_\theta \log \pi_\theta(a \mid s) \cdot A^{\pi_\theta}(s, a)]$

Questa struttura è il cuore dei metodi actor-critic: un componente actor apprende la politica π_θ usando il segnale di vantaggio come guida; un componente critic apprende la funzione di valore $V^{\pi_\theta}_\varphi$ con parametri φ, per stimare il vantaggio. I due componenti si co-addestrano, con il critic che fornisce all'actor un segnale di gradiente a varianza

Silvano Salvador

ridotta, e l'actor che genera le traiettorie su cui il critic aggiorna la propria stima del valore.

IV. La stima del vantaggio generalizzata (GAE)

La stima pratica della funzione di vantaggio $A^\pi(s_t, a_t)$ richiede una scelta tra bias e varianza. Stime basate su un solo passo temporale — l'errore TD a un passo, $\delta_t = r_t + \gamma V^\pi(s_{t+1}) - V^\pi(s_t)$ — hanno varianza bassa ma sono distorte (dipendono dalla qualità dell'approssimazione del critic V^π_φ). Stime basate sull'intera traiettoria Monte Carlo — $G_t - V^\pi(s_t)$ — sono non distorte ma ad alta varianza. La Generalized Advantage Estimation (GAE) di Schulman e colleghi (2016) interpoltra tra questi due estremi con un parametro $\lambda \in [0, 1]$:

$$\hat{A}_t^{GAE(\gamma,\lambda)} = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}$$

dove $\delta_{t+l} = r_{t+l} + \gamma V^\pi_\varphi(s_{t+l+1}) - V^\pi_\varphi(s_{t+l})$ è l'errore TD all'passo t+l.

Per verificare questa formula, si espande la somma dei λ-ritorni. Si definisce $G_t^{(k)} = \sum_{l=0}^{k-1} \gamma^l r_{t+l} + \gamma^k V^\pi_\varphi(s_{t+k})$ il k-step return. La GAE si ottiene come media geometrica pesata delle stime di vantaggio a k passi:

$$\hat{A}_t^{GAE} = (1 - \lambda) \sum_{k=1}^{\infty} \lambda^{k-1} (G_t^{(k)} - V^\pi_\varphi(s_t))$$

Si noti che per $\lambda = 0$ si ottiene $\hat{A}_t = \delta_t = r_t + \gamma V(s_{t+1}) - V(s_t)$, la stima a un passo (bassa varianza, alto bias); per $\lambda = 1$ si ottiene $\hat{A}_t = \sum_{l=0}^{\infty} \gamma^l r_{t+l} - V(s_t)$, il ritorno Monte Carlo scontato (bassa distorsione, alta varianza). Il parametro λ controlla il compromesso bias-varianza.

Silvano Salvador

La formula ricorsiva per il calcolo efficiente di GAE è:

$$\hat{A}_t^{GAE} = \delta_t + (\gamma\lambda)\,\hat{A}_{t+1}^{GAE}$$

che si ottiene dalla ricorsione $\sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l} = \delta_t + \gamma\lambda \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+1+l}$. Questa formula ricorsiva permette il calcolo di tutti i vantaggi GAE di una traiettoria di lunghezza T in O(T) operazioni, scorrendo la traiettoria all'indietro nel tempo. La struttura ricorsiva è analoga alla retropropagazione del segnale di errore nelle reti neurali, e questa non è una coincidenza: entrambi i calcoli corrispondono alla propagazione all'indietro di un segnale di critica attraverso una sequenza temporale di operazioni.

V. Trust Region Policy Optimization: la stabilità come requisito geometrico

Un difetto comune a tutti i metodi di policy gradient è l'instabilità dell'ottimizzazione: aggiornamenti troppo grandi dei parametri θ possono portare la politica in regioni dello spazio dei parametri dove essa si comporta molto peggio della politica corrente. Poiché le nuove traiettorie sono raccolte sotto la politica aggiornata, un degrado della politica si autoalimenta: le traiettorie peggiorano, i segnali di gradiente diventano meno informativi, e il processo di apprendimento può deteriorarsi in modo catastrofico e difficile da correggere. L'entità degli aggiornamenti è quindi un parametro critico, e la sua scelta corretta è un prerequisito per la convergenza stabile.

Il Trust Region Policy Optimization (TRPO) di Schulman, Levine, Moritz, Jordan e Abbeel (2015) risolve questo problema con rigore teorico, imponendo un vincolo esplicito sulla dimensione

Silvano Salvador

dell'aggiornamento della politica, misurata in termini di divergenza KL media tra la politica vecchia e quella nuova:

$$\max_{\theta} \mathbb{E}_{s \sim \varrho^{\pi_{\theta_{old}}},\, a \sim \pi_{\theta_{old}}(\cdot|s)} [r_t(\theta) \cdot \hat{A}(s, a)]$$
soggetto a: $\mathbb{E}_s [D_{KL}(\pi_{\theta_{old}}(\cdot \mid s) \,\|\, \pi_\theta(\cdot \mid s))] \leq \delta$

dove $r_t(\theta) = \pi_\theta(a \mid s) / \pi_{\theta_{old}}(a \mid s)$ è il rapporto di importanza (importance ratio) — il rapporto tra la probabilità dell'azione sotto la politica nuova e quella sotto la politica vecchia — e δ è il raggio della regione di fiducia. Il termine $r_t(\theta) \cdot \hat{A}(s, a)$ è l'obiettivo surrogato (surrogate objective): esso approssima il guadagno effettivo della politica π_θ rispetto a $\pi_{\theta_{old}}$ usando traiettorie raccolte sotto $\pi_{\theta_{old}}$ (importance sampling), senza dover raccogliere nuove traiettorie per ogni candidato aggiornamento.

La giustificazione teorica di TRPO è il lemma di politica migliorata di Kakade e Langford (2002), che fornisce un limite inferiore al guadagno reale $J(\theta) - J(\theta_{old})$ in termini dell'obiettivo surrogato e della divergenza KL:

$$J(\theta) - J(\theta_{old}) \geq L_{\theta_{old}}(\theta) - C \cdot D_{KL}^{max}(\pi_{\theta_{old}} \,\|\, \pi_\theta)$$

dove $L_{\theta_{old}}(\theta) = \mathbb{E}_{s \sim d^{\pi_{\theta_{old}}},\, a \sim \pi_{\theta_{old}}} [r_t(\theta) \cdot \hat{A}(s, a)]$ è l'obiettivo surrogato, $D_{KL}^{max}(\pi_{\theta_{old}} \,\|\, \pi_\theta) = \max_s D_{KL}(\pi_{\theta_{old}}(\cdot \mid s) \,\|\, \pi_\theta(\cdot \mid s))$ è la divergenza KL massima sugli stati, e $C = 4 \varepsilon_{max} \gamma / (1 - \gamma)^2$ con $\varepsilon_{max} = \max_{s,a} |\hat{A}(s, a)|$ è una costante.

Il limite inferiore è positivo quando $L_{\theta_{old}}(\theta)$ è sufficientemente maggiore del termine di penalizzazione $C \cdot D_{KL}^{max}$. Massimizzando questo limite inferiore — cioè trovando θ che massimizza $L_{\theta_{old}}(\theta)$ sotto il vincolo sulla divergenza KL — si garantisce che la nuova politica non sia peggiore della vecchia.

Silvano Salvador

L'algoritmo TRPO implementa questo ottimizzazione risolvendo un problema vincolato con il metodo di Newton approssimato: l'obiettivo è linearizzato e il vincolo è approssimato al secondo ordine, producendo un'equazione di Newton la cui soluzione fornisce la direzione di aggiornamento ottimale; poi si esegue un backtracking per trovare la dimensione del passo che rispetta il vincolo sulla divergenza KL.

La direzione di Newton richiede il calcolo del gradiente della politica e dell'Hessiana approssimata della divergenza KL. Quest'ultima coincide, al primo ordine, con la matrice di informazione di Fisher della politica — una connessione profonda che emergerà nella prossima sezione sull'ottimizzazione naturale. Il calcolo esatto del prodotto di questa matrice per un vettore è computazionalmente costoso ma fattibile attraverso il metodo del gradiente coniugato applicato all'equazione $Fx = g$, dove F è la matrice di Fisher e g è il gradiente dell'obiettivo.

VI. Proximal Policy Optimization: semplicità e robustezza pratica

TRPO è teoricamente ben fondato ma computazionalmente oneroso: richiede la soluzione di un problema di ottimizzazione vincolato a ogni aggiornamento, con calcolo del gradiente naturale attraverso il gradiente coniugato e backtracking per rispettare il vincolo KL. Il Proximal Policy Optimization (PPO) di Schulman, Wolski, Dhariwal, Radford e Klimov (2017) sostituisce il vincolo esplicito sulla divergenza KL con una modifica dell'obiettivo surrogato — il clipping del rapporto di importanza — che realizza indirettamente la stessa restrizione sulla dimensione dell'aggiornamento, in modo molto più semplice da implementare e da ottimizzare.

L'obiettivo PPO-clip è:

Silvano Salvador

$$L^{CLIP}(\theta) = \mathbb{E}_t [\min(r_t(\theta) \cdot \hat{A}_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon) \cdot \hat{A}_t)]$$

dove $r_t(\theta) = \pi_\theta(a_t \mid s_t) / \pi_{\theta_{old}}(a_t \mid s_t)$ è il rapporto di importanza, $\hat{A}_t$ è la stima del vantaggio, e ε è un iperparametro tipicamente scelto nell'intervallo [0.1, 0.3].

La comprensione intuitiva di questo obiettivo si ottiene esplicitando i due casi in base al segno del vantaggio. Quando $\hat{A}_t \geq 0$ — l'azione intrapresa era migliore della media, e si vuole aumentare la sua probabilità — l'obiettivo è $\min(r_t(\theta), 1+\varepsilon) \cdot \hat{A}_t$. Aumentare $r_t(\theta)$ aumenta l'obiettivo, ma il clipping a $1+\varepsilon$ impedisce aumenti eccessivi del rapporto di probabilità: oltre la soglia $1+\varepsilon$, ulteriori aumenti non producono guadagno nell'obiettivo. Quando $\hat{A}_t < 0$ — l'azione era peggiore della media, e si vuole diminuire la sua probabilità — l'obiettivo è $\max(r_t(\theta), 1-\varepsilon) \cdot \hat{A}_t$. Diminuire $r_t(\theta)$ verso zero rende l'azione meno probabile, ma il clipping a $1-\varepsilon$ impedisce che $r_t(\theta)$ scenda troppo: la politica non può cambiare troppo bruscamente in nessuna direzione.

Il meccanismo del clipping implementa una forma di penalizzazione conservativa: ogni volta che l'aggiornamento tenderebbe a spingere il rapporto $r_t(\theta)$ fuori dall'intervallo $[1-\varepsilon, 1+\varepsilon]$, l'obiettivo diventa piatto nella direzione dell'aggiornamento, e il gradiente si annulla. Questo scoraggia aggiornamenti che porterebbero la politica troppo lontano da quella vecchia, in modo analogo al vincolo KL di TRPO ma senza il costo computazionale della sua applicazione esplicita.

In termini di complessità implementativa, PPO riduce il problema a una semplice ottimizzazione a gradiente stocastico con una funzione di perdita modificata, computabile con una singola forward-backward pass su un mini-batch di transizioni. Questo ne ha fatto l'algoritmo di policy gradient più usato nella pratica: è il fondamento dell'addestramento di molti sistemi robotici, dei sistemi di gioco di alta performance, e — più

Silvano Salvador

rilevante per il contesto di questo saggio — dell'allineamento dei grandi modelli linguistici mediante RLHF, come si discuterà nella Parte Quarta.

L'obiettivo PPO completo combina il termine di clipping con una stima del valore e un termine di entropia:

$$L^{PPO}(\theta, \varphi) = \mathbb{E}_t [L^{CLIP}_t(\theta) - c_1 \cdot (V_\varphi(s_t) - V_t^{target})^2 + c_2 \cdot H(\pi_\theta(\cdot \mid s_t))]$$

dove il secondo termine allena il critic V_φ a stimare il valore degli stati (indispensabile per calcolare $\hat{A}_t$), il terzo termine aggiunge un bonus di entropia H che incoraggia l'esplorazione impedendo alla politica di collassare prematuramente su un sottoinsieme di azioni, e c_1, c_2 sono coefficienti che bilanciano i tre contributi.

VII. Il gradiente naturale e la geometria di Fisher

Per comprendere pienamente la struttura di TRPO e PPO, è utile sviluppare la connessione con la geometria dell'informazione che ne costituisce il fondamento teorico. Il gradiente euclideo ordinario $\nabla_\theta J(\theta)$ assume una metrica euclidea sullo spazio dei parametri: tutti i passi $\|\Delta\theta\| = \varepsilon$ sono trattati come ugualmente grandi. Ma questa metrica non è quella naturale per lo spazio delle politiche: due politiche con valori numerici dei parametri molto diversi ma comportamentalmente identiche sono alla distanza zero in termini di comportamento, mentre due politiche con parametri numericamente vicini ma comportamentalmente distanti si trovano lontane in termini di effetti sull'ambiente.

La geometria naturale sullo spazio delle politiche è quella indotta dalla divergenza KL, ovvero la metrica di informazione di Fisher. La matrice di Fisher della politica π_θ è:

Silvano Salvador

$$F(\theta) = \mathbb{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta(\cdot|s)} [\nabla_\theta \log \pi_\theta(a \mid s) \cdot (\nabla_\theta \log \pi_\theta(a \mid s))^T]$$

Questa matrice ha una doppia interpretazione. Come matrice di covarianza delle score functions della politica, essa misura quanto rapidamente cambia la distribuzione delle azioni al variare dei parametri in diverse direzioni. Come metrica riemanniana sullo spazio dei parametri, essa assegna a ogni vettore tangente $\Delta\theta$ una lunghezza riemanniana $\|\Delta\theta\|_F^2 = \Delta\theta^T F(\theta) \Delta\theta$ che misura la variazione comportamentale della politica nella direzione $\Delta\theta$ — una variazione che coincide, al secondo ordine, con la divergenza KL:

$$D_{KL}(\pi_\theta \| \pi_{\theta+\Delta\theta}) = (1/2) \Delta\theta^T F(\theta) \Delta\theta + O(\|\Delta\theta\|^3)$$

Il gradiente naturale di Amari (1998), applicato al reinforcement learning da Kakade (2002), è la direzione di massima ascesa di $J(\theta)$ nella metrica di Fisher:

$$\tilde{\nabla}_\theta J(\theta) = F(\theta)^{-1} \nabla_\theta J(\theta)$$

Un passo nella direzione del gradiente naturale produce il massimo aumento di $J(\theta)$ per unità di variazione comportamentale della politica — cioè, il massimo aumento di performance per unità di divergenza KL dall'aggiornamento. La connessione con TRPO è precisa e fondamentale: l'aggiornamento di TRPO con regione di fiducia nella metrica KL è, al primo ordine, equivalente a un passo del gradiente naturale. La dimensione del passo è determinata dal raggio δ della regione di fiducia, che corrisponde alla divergenza KL massima ammissibile.

La matrice di Fisher $F(\theta)$ ha anche un'interpretazione come curvatura della superficie di log-verosimiglianza della politica: essa misura la curvatura dello spazio dei parametri nella metrica dell'informazione, rendendone le isocontorni delle sfere nella metrica di Fisher anziché

Silvano Salvador

sfere euclidee. Ottimizzare nella metrica di Fisher significa muoversi sulle geodesiche della varietà statistica dei parametri della politica — le curve di minima lunghezza nella geometria dell'informazione — il che garantisce che ogni passo di aggiornamento produca un cambiamento comportamentale controllato e significativo, invariante rispetto alla parametrizzazione scelta.

Questa invarianza rispetto alla parametrizzazione è una proprietà fondamentale che il gradiente euclideo ordinario non possiede. Se si reparametrizza la politica con una funzione biunivoca $\theta' = f(\theta)$, il gradiente ordinario cambia covariante (e il punto di massimo può sembrare spostarsi), mentre il gradiente naturale rimane invariante: indica la stessa direzione di miglioramento comportamentale indipendentemente dalla coordinatizzazione dello spazio dei parametri. Questa proprietà è essenziale per costruire algoritmi di ottimizzazione che siano robusti rispetto alle scelte di architettura e parametrizzazione della politica.

VIII. Actor-Critic asincrono e le architetture scalabili

I metodi actor-critic moderni — nei quali un componente actor aggiorna la politica usando il segnale di vantaggio stimato da un critic che apprende la funzione di valore — riescono a sfruttare la potenza computazionale distribuita attraverso architetture parallele. L'Asynchronous Advantage Actor-Critic (A3C) di Mnih, Badia, Mirza, Graves, Harley, Lillicrap, Silver e Kavukcuoglu (2016) esegue molti thread paralleli, ciascuno con la propria copia dell'ambiente e con accesso ai parametri globali condivisi, e aggiorna questi parametri in modo asincrono man mano che ciascun thread accumula esperienza da

Silvano Salvador

traiettorie brevi. La diversità delle esperienze raccolte da ambienti paralleli sostituisce l'experience replay come meccanismo di decorrelazione: invece di campionare da un buffer storico, si raccolgono esperienze simultanee da istanze diverse dell'ambiente.

La versione sincrona A2C raccoglie traiettorie da N agenti in parallelo e le usa tutte per un singolo aggiornamento sincrono, garantendo che tutte le esperienze siano raccolte sotto la medesima versione della politica. Questa proprietà — che tutte le esperienze condividano la stessa politica di comportamento — semplifica l'analisi teorica e riduce la varianza del gradiente stimato rispetto alla versione asincrona, al costo di un utilizzo meno efficiente del parallelismo computazionale.

La razionalità in conflitto: teoria dei giochi e intelligenza multi-agente

---

I. Oltre l'agente singolo: la complessità dell'interazione strategica

L'intera architettura matematica sviluppata nelle parti precedenti presuppone un agente singolo che interagisce con un ambiente stazionario: un ambiente le cui proprietà statistiche non dipendono dalla politica dell'agente stesso. Questa assunzione è soddisfatta in molti contesti di interesse — giochi contro simulatori deterministici, sistemi di controllo di robot fisici, ottimizzazione di strategie in ambienti naturali — ma fallisce non appena l'agente interagisce con altri agenti che a loro volta apprendono e adattano le proprie strategie. In questo caso

Silvano Salvador

l'ambiente è non stazionario per definizione: le transizioni e le ricompense dipendono dalle politiche di tutti gli agenti, che si modificano nel tempo violando sistematicamente le condizioni di convergenza di tutti gli algoritmi di reinforcement learning per agente singolo.

La teoria dei giochi — fondata da John von Neumann e Oskar Morgenstern nel 1944 con la Theory of Games and Economic Behavior, e poi estesa da John Nash con i contributi del 1950 e 1951 che gli valsero il Premio Nobel nel 1994 — è il quadro matematico appropriato per l'analisi delle interazioni strategiche tra agenti razionali. La sua rilevanza per l'intelligenza artificiale è duplice. Sul piano tecnico, fornisce il framework per l'addestramento di agenti in ambienti multi-agente: il self-play che ha prodotto AlphaZero, i sistemi di negoziazione automatica, le aste algoritmiche sono tutti istanze di questo framework. Sul piano filosofico — e questo è il punto di connessione con il problema dell'allineamento sviluppato nella parte successiva — la teoria dei giochi rivela strutture di instabilità e incompatibilità tra razionalità individuale e razionalità collettiva che hanno implicazioni profonde per la progettazione di sistemi di AI cooperativi con gli esseri umani.

II. Struttura formale di un gioco e il concetto di equilibrio

Un gioco in forma normale è una tripla (N, $\{\mathcal{A}_i\}_{i \in N}$, $\{u_i\}_{i \in N}$) dove N = {1, ..., n} è l'insieme dei giocatori, $\mathcal{A}_i$ è lo spazio delle azioni del giocatore i, e $u_i : \prod_{j \in N} \mathcal{A}_j \to \mathbb{R}$ è la funzione di utilità del giocatore i, che assegna un payoff reale a ogni profilo di azioni $(a_1, ..., a_n) \in \prod_j \mathcal{A}_j$. Una strategia mista σ_i è una distribuzione di probabilità su

Silvano Salvador

$\mathcal{A}_i$ — una scelta randomizzata tra le azioni pure. Il payoff atteso del giocatore i sotto un profilo di strategie miste $\sigma = (\sigma_1, ..., \sigma)$ è

$u_i(\sigma) = \sum_{a \in \prod_j \mathcal{A}_j} (\prod_j \sigma(a)) \cdot u_i(a)$

che è multilineare nelle componenti di σ: lineare in σ_i per ogni σ_{-i} fissato. Questa multilinearità è la proprietà fondamentale che garantisce la convessità delle migliori risposte e rende applicabile il teorema di Kakutani.

Il concetto soluzione centrale della teoria dei giochi è l'equilibrio di Nash: un profilo di strategie $\sigma^* = (\sigma_1^*, ..., \sigma^*)$ tale che nessun giocatore ha incentivo a deviare unilateralmente dalla propria strategia data quella degli altri:

$u_i(\sigma_i^*, \sigma_{-i}) \geq u_i(\sigma_i, \sigma_{-i})$ per ogni $i \in N$ e ogni $\sigma_i \in \Delta(\mathcal{A}_i)$

dove $\sigma_{-i}^* = (\sigma_1^*, ..., \sigma_{i-1}, \sigma_{i+1}, ..., \sigma^*)$ indica le strategie di tutti i giocatori eccetto i, e $\Delta(\mathcal{A}_i)$ è il simplesso delle distribuzioni di probabilità su $\mathcal{A}_i$. La condizione afferma che σ_i^* è la migliore risposta di i alla strategia degli avversari: nessuna strategia alternativa di i produce un payoff atteso più alto, dati i comportamenti fissi degli altri.

La formalizzazione del concetto di migliore risposta è la seguente. La corrispondenza di migliore risposta del giocatore i è

$BR_i(\sigma_{-i}) = \text{argmax}_{\sigma_i \in \Delta(\mathcal{A}_i)} u_i(\sigma_i, \sigma_{-i})$

Per multilinearità di u_i in σ_i, il massimo su $\Delta(\mathcal{A}_i)$ è attinto sulle strategie miste il cui supporto è l'insieme delle azioni pure che massimizzano $u_i(\cdot, \sigma_{-i})$:

$BR_i(\sigma_{-i}) = \text{conv}\{a_i \in \mathcal{A}_i : u_i(a_i, \sigma_{-i}) = \max_{b_i \in \mathcal{A}_i} u_i(b_i, \sigma_{-i})\}$

Silvano Salvador

che è un sottoinsieme convesso non vuoto di $\Delta(\mathcal{A}_i)$. Un equilibrio di Nash è un punto fisso della corrispondenza globale di migliore risposta $BR(\sigma) = \prod i\ BR_i(\sigma\{-i\})$.

III. Il teorema di esistenza di Nash: Kakutani e la topologia del punto fisso

Il teorema di esistenza di Nash (1950) afferma che ogni gioco finito — con n giocatori e spazi di azione finiti — ammette almeno un equilibrio di Nash in strategie miste. La dimostrazione si basa sul teorema del punto fisso di Kakutani per corrispondenze, un'estensione del teorema di Brouwer per funzioni continue.

Teorema (Kakutani, 1941). Sia X un sottoinsieme compatto e convesso di $\mathbb{R}^m$, e sia $\varphi : X \to 2^X$ una corrispondenza tale che: (i) $\varphi(x) \neq \varnothing$ per ogni $x \in X$; (ii) $\varphi(x)$ è convessa per ogni $x \in X$; (iii) il grafo di φ, $Gr(\varphi) = \{(x,y) : y \in \varphi(x)\}$, è un insieme chiuso. Allora esiste $x^* \in X$ tale che $x^* \in \varphi(x^*)$.

Per applicare questo teorema alla dimostrazione dell'equilibrio di Nash, si verifica che la corrispondenza globale di migliore risposta $BR : \Delta \to 2^\Delta$ soddisfi le ipotesi del teorema di Kakutani, dove $\Delta = \prod_i \Delta(\mathcal{A}_i)$ è il prodotto dei simplessi delle strategie miste.

Il dominio Δ è compatto (prodotto finito di simplessi compatti) e convesso (prodotto finito di insiemi convessi). I valori $BR(\sigma)$ sono non vuoti per ogni $\sigma \in \Delta$: poiché $u_i(\cdot, \sigma_{-i})$ è continua (lineare, quindi continua) su Δ_i compatto, il massimo è attinto dal teorema di Weierstrass. I valori $BR(\sigma)$ sono convessi per ogni σ: $BR_i(\sigma_{-i})$ è il simplesso sulle azioni pure ottimali, un simplesso facciale di Δ_i, che è convesso; il prodotto $BR(\sigma) = \prod i\ BR_i(\sigma\{-i\})$ è convesso come prodotto di insiemi

Silvano Salvador

convessi. La chiusura del grafo di BR si verifica usando la continuità di u_i: se $(\sigma^{(k)}, \tau^{(k)}) \to (\sigma^*, \tau^*)$ con $\tau^{(k)} \in BR(\sigma^{(k)})$, allora per ogni deviazione alternativa $\tilde{\sigma}_i$ e per ogni k:

$u_i(\tau_i^{(k)}, \sigma_{-i}^{(k)}) \geq u_i(\tilde{\sigma}_i, \sigma_{-i}^{(k)})$

Passando al limite per continuità di u_i: $u_i(\tau_i^*, \sigma_{-i}) \geq u_i(\tilde{\sigma}_i, \sigma_{-i})$ per ogni $\tilde{\sigma}_i$, dunque $\tau^* \in BR(\sigma^*)$ e il grafo è chiuso. Per il teorema di Kakutani, esiste $\sigma^* \in BR(\sigma^*)$ — esattamente la definizione di equilibrio di Nash. □

IV. Il teorema minimax: dualità e valore del gioco a somma zero

Nei giochi a somma zero a due giocatori — dove $u_1(a) + u_2(a) = 0$ per ogni profilo di azioni a, cosicché ciò che guadagna P_1 è esattamente ciò che perde P_2 — la struttura degli equilibri è particolarmente ricca. Il fondamentale teorema di von Neumann (1928) afferma che il valore del gioco — il payoff atteso in equilibrio — è ben definito e unico, e può essere calcolato come soluzione di un programma lineare.

Sia $M \in \mathbb{R}^{m \times n}$ la matrice dei payoff, dove M_{ij} è il payoff di P_1 (che P_2 corrispettivamente perde) quando P_1 gioca l'azione pura $i \in \{1,...,m\}$ e P_2 gioca l'azione pura $j \in \{1,...,n\}$. Il payoff atteso di P_1 sotto strategie miste $(p, q) \in \Delta_m \times \Delta_n$ è $p^T M q$.

Teorema minimax di von Neumann:

$\max_{p \in \Delta_m} \min_{q \in \Delta_n} p^T M q = \min_{q \in \Delta_n} \max_{p \in \Delta_m} p^T M q =: V^*$

Il valore V^* è il valore del gioco: P_1 può garantirsi almeno V^* indipendentemente dalla strategia di P_2 (giocando la strategia minimax

Silvano Salvador

p^*), e P_2 può garantire a P_1 al massimo V^* (giocando la strategia minimax q^*).

La dimostrazione via dualità LP procede come segue. Il problema di P_1 — trovare la strategia mista che massimizza il peggior caso contro P_2 — è:

(P_1) $\max V$ s.t. $\sum_i p_i M\{ij\} \geq V \ \forall j$, $\sum_i p_i = 1$, $p \geq 0$, V libero

Sostituendo $V' = V + K$ per K tale che $V' > 0$ e $p'_i = p_i / V'$, si ottiene il programma lineare equivalente:

(LP_1) $\min \sum_i p'_i$ s.t. $\sum_i p'_i M\{ij\} \geq 1 \ \forall j$, $p' \geq 0$

Il duale LP di (LP_1) è:

(LP_2) $\max \sum_j q' \ \text{s.t.} \ \sum_j q' \ M\{ij\} \leq 1 \ \forall i$, $q' \geq 0$

che corrisponde al problema di minimizzazione di P_2. Per la dualità forte della programmazione lineare — che vale perché (LP_1) è ammissibile e limitato — i valori ottimi primale e duale coincidono, il che implica $\max_{p} \min_{q} p^T M q = \min_{q} \max_{p} p^T M q = V^*$. □

La dimostrazione alternativa via separazione di insiemi convessi — che usa il teorema di separazione degli iperiani per disgiunzione di due insiemi convessi — è equivalente e rivela la struttura geometrica sottostante: il valore del gioco è il punto di sella della funzione di payoff, il punto di incontro tra la curva di massimo di P_1 e quella di minimo di P_2.

V. Il dilemma del prigioniero: razionalità individuale e collettiva

Silvano Salvador

Il dilemma del prigioniero è forse il gioco più studiato nella storia della teoria dei giochi, e la sua importanza trascende ampiamente la teoria dei giochi stessa, toccando questioni fondamentali di filosofia morale, economia politica, biologia evolutiva e — come si vedrà — ingegneria dei sistemi di AI allineati. Due giocatori scelgono simultaneamente tra cooperare (C) e deferrire (D). Il payoff in forme matriciale è tale che i payoff di entrambi i giocatori soddisfino: deferire domina strettamente cooperare per ciascun giocatore (qualunque sia la scelta dell'avversario, il giocatore ottiene un payoff maggiore deferendo), ma il risultato (C, C) è paretianamente superiore al risultato (D, D).

Il profilo specifico tipicamente usato ha i seguenti payoff per (P_1, P_2): (C,C)→(3,3), (C,D)→(0,5), (D,C)→(5,0), (D,D)→(1,1). L'unico equilibrio di Nash è (D, D) con payoff (1,1): deferire è la strategia strettamente dominante per entrambi i giocatori. Tuttavia l'esito (C, C) con payoff (3, 3) è paretianamente superiore: entrambi i giocatori starebbero meglio cooperando.

Il dilemma del prigioniero esibisce un conflitto strutturale tra razionalità individuale e razionalità collettiva: l'equilibrio di Nash — il risultato della razionalità individuale disaccoppiata — è sub-ottimale dal punto di vista dell'utilità aggregata. Questa struttura non è un artefatto della forma specifica del gioco: è la manifestazione dell'incompatibilità più generale tra ottimizzazione individuale e ottimizzazione collettiva che il teorema di impossibilità di Arrow formalizza in un contesto ancora più generale, e che la legge di Goodhart istanzia nel contesto dell'ottimizzazione di metriche proxy. Per l'intelligenza artificiale, il dilemma del prigioniero descrive con precisione formale la struttura di molti problemi reali: la gestione dei beni comuni digitali, la negoziazione tra agenti AI con funzioni obiettivo diverse, il coordinamento tra sistemi autonomi che perseguono obiettivi parzialmente contrapposti.

Silvano Salvador

Nel dilemma del prigioniero ripetuto — dove lo stesso gioco viene giocato un numero T di volte tra gli stessi giocatori, con ciascun giocatore che osserva le mosse passate dell'avversario — la struttura cambia radicalmente. Il Folk theorem della teoria dei giochi ripetuti afferma che, se il gioco viene ripetuto infinitamente (o con probabilità abbastanza alta di continuazione), ogni profilo di payoff individualmente razionale e paretianamente ammissibile è sostenibile come equilibrio di Nash in strategie trigger. Formalmente, con il fattore di sconto γ, la cooperazione attraverso le strategie grim trigger — coopera finché l'avversario coopera, defetta permanentemente al primo tradimento — è un equilibrio se e soltanto se:

$$u_C / (1 - \gamma) \geq u_D + \gamma \cdot u_DD / (1 - \gamma)$$

dove u_C è il payoff dalla cooperazione reciproca, u_D è il payoff dalla deviazione unilaterale, e u_DD è il payoff dalla defazione reciproca permanente. Risolvendo per γ:

$$\gamma \geq (u_D - u_C) / (u_D - u_DD)$$

Per i payoff tipici del dilemma del prigioniero (u_C = 3, u_D = 5, u_DD = 1), la soglia è $\gamma \geq (5-3)/(5-1) = 1/2$: per qualsiasi fattore di sconto superiore a 1/2, la cooperazione è razionalmente sostenibile come equilibrio. La cooperazione è razionale nell'orizzonte lungo perché il guadagno immediato dalla defezione — u_D − u_C = 2 — è inferiore alla perdita futura attualizzata dalla punizione — $\gamma(u_C - u_DD)/(1-\gamma)$.

VI. Le dinamiche di replica e il gradiente naturale evolutivo

Il collegamento tra teoria dei giochi e biologia evolutiva è realizzato attraverso le dinamiche di replica (replicator dynamics) di Taylor e Jonker

Silvano Salvador

(1978), che descrivono l'evoluzione della composizione di una popolazione che gioca un gioco simmetrico sotto pressione di selezione naturale. Sia $x = (x_1, ..., x)$ il vettore delle frequenze degli n tipi strategici in una popolazione, con $\sum_i x_i = 1$. Il fitness atteso del tipo i quando la composizione della popolazione è x è $f_i(x) = (Mx)_i = \sum_j M_{ij} x$, dove M è la matrice dei payoff del gioco simmetrico. Il fitness medio della popolazione è $\bar{f}(x) = x^T M x = \sum_i x_i f_i(x)$.

Le replicator dynamics sono il sistema di equazioni differenziali ordinarie:

$\dot{x}_i = x_i (f_i(x) - \bar{f}(x))$ per ogni $i = 1, ..., n$

Il simplesso Δ_{n-1} è invariante per queste dinamiche: $\sum_i \dot{x}_i = \sum_i x_i(f_i - \bar{f}) = \sum_i x_i f_i - \bar{f} \sum_i x_i = \bar{f} - \bar{f} \cdot 1 = 0$. Il tipo i cresce in frequenza quando il suo fitness supera la media, e decresce altrimenti — un'implementazione demografica del principio darwiniano della selezione per adattamento relativo.

La connessione con il gradiente naturale — tra le più profonde nell'intera teoria dell'ottimizzazione computazionale — è la seguente. Le replicator dynamics possono essere scritte nella forma:

$\dot{x} = G(x)^{-1} \nabla_x \bar{f}(x) | \{\Delta\}$

dove G(x) è la matrice di Fisher della distribuzione categoriale con parametro x, e $\nabla_x \bar{f}|_\Delta$ è il gradiente del fitness medio ristretto al piano tangente del simplesso. Questo risultato afferma che le dinamiche evolutive della selezione naturale sono esattamente le dinamiche del gradiente naturale dell'adattamento medio — il gradiente adattato alla geometria di Fisher dello spazio dei tipi.

La verifica richiede di calcolare esplicitamente il gradiente naturale. La distribuzione categoriale $p(i; x) = x_i$ ha score function $\partial \log x_i / \partial x =$

Silvano Salvador

$\delta_i\square/x_i$. La matrice di Fisher sul piano tangente del simplesso (con il vincolo $\sum x_i = 1$) ha elementi $G_{ij}(x) = \delta_i\square/x_i$ rispetto alle coordinate affini libere. Il gradiente del fitness medio ristretto al simplesso è $[\nabla_\Delta f]_i = 2(Mx)_i - 2f(x) = 2(f_i(x) - f(x))$ per simmetria di M (usando $\partial f/\partial x_i = 2(Mx)_i$). Il gradiente naturale è $G(x)^{-1}\ \nabla_\Delta f = \mathrm{diag}(x_i) \cdot (2(f - f\,1)) = 2x_i(f_i - f)$, proporzionale al membro destro delle replicator dynamics. □

Questa identità matematica ha un'implicazione filosofica notevole he di è cercato di sviluppare in vari modi: l'ottimizzazione evolutiva e l'ottimizzazione della politica nel reinforcement learning con gradiente naturale sono, a un livello di astrazione appropriato, la stessa struttura computazionale. La selezione naturale su popolazioni di strategie è il reinforcement learning implementato dalla demografia; il reinforcement learning con gradiente naturale è la selezione naturale implementata dall'algebra differenziale. Il cervello — prodotto dell'evoluzione, che implementa meccanismi di rinforzo dopaminergico corrispondenti al TD error — è il punto di incontro biologico di queste due strutture matematicamente equivalenti.

Il teorema fondamentale della selezione naturale di Fisher (1930), visto come proprietà delle replicator dynamics, afferma che il fitness medio f(x) non decresce mai lungo le traiettorie:

$$d\,f/dt = \mathrm{Var}_x[f] = \sum_i x_i\,(f_i - f)^2 \geq 0$$

La dimostrazione è immediata: $d\,f/dt = \sum_i \dot{x}_i\ f_i = \sum_i x_i(f_i - f)f_i = \mathbb{E}_x[f^2] - (\mathbb{E}_x[f])^2 = \mathrm{Var}_x[f] \geq 0$. Il fitness medio non decresce mai, e si mantiene costante soltanto quando la varianza del fitness è nulla — cioè in equilibrio, quando tutti i tipi presenti hanno lo stesso fitness. La selezione naturale è un processo di massimizzazione del fitness medio rispetto alla geometria di Fisher del simplesso evolutivo.

Silvano Salvador

VII. Il multi-agent reinforcement learning e la non-stazionarietà

Il multi-agent reinforcement learning (MARL) è il campo che studia il problema dell'apprendimento simultaneo di strategie ottimali da parte di più agenti che interagiscono in un ambiente comune. La difficoltà fondamentale rispetto al reinforcement learning per agente singolo è la non-stazionarietà: quando tutti gli agenti apprendono simultaneamente, il processo stocastico generato dall'interazione non è una catena di Markov stazionaria, perché la distribuzione delle transizioni dipende dalle politiche di tutti gli agenti, che cambiano nel tempo. Qualsiasi algoritmo di reinforcement learning per agente singolo presuppone che l'ambiente sia stazionario — presupposto sistematicamente violato.

Formalmente, l'interazione multi-agente è modellata come un Markov game (o stochastic game), introdotto da Shapley nel 1953: una sextupla $(N, \mathcal{S}, \{\mathcal{A}_i\}_{i \in N}, P, \{R_i\}_{i \in N}, \gamma)$ dove N è il numero di agenti, $\mathcal{S}$ lo spazio degli stati condiviso, $\mathcal{A}_i$ lo spazio delle azioni dell'agente i, $P(s' \mid s, a_1, ..., a)$ la funzione di transizione che dipende dalle azioni di tutti gli agenti, $R_i(s, a_1, ..., a)$ la funzione di ricompensa dell'agente i, e γ il fattore di sconto. L'MDP per agente singolo è il caso speciale $n = 1$.

Il concetto soluzione per un Markov game è il Nash equilibrium di politiche markoviane: un profilo $(\pi_1^*, ..., \pi^*)$ tale che per ogni agente i e ogni stato s:

$$V_i^{\pi^*}(s) \geq V_i^{\pi_i, \pi^*_{-i}}(s)$$

dove $V_i^{\pi}(s)$ è il valore dell'agente i nello stato s sotto il profilo π, e (π_i, π^*_{-i}) denota il profilo ottenuto sostituendo la politica di i con π_i e mantenendo fisse quelle degli altri. L'esistenza di un tale equilibrio è garantita per Markov game finiti stazionari, ma convergere a esso

Silvano Salvador

attraverso l'apprendimento simultaneo è significativamente più difficile: nessun algoritmo di MARL ha le stesse garanzie di convergenza del Q-learning per agente singolo nel caso generale.

Il problema del credit assignment in ambienti multi-agente cooperativi — come attribuire la ricompensa collettiva alle azioni dei singoli agenti — è una difficoltà aggiuntiva senza analogo nel caso singolo. Se N agenti cooperano per raggiungere un obiettivo condiviso e ricevono una ricompensa comune r al termine, ciascun agente riceve lo stesso segnale di rinforzo, senza informazione su quanto la propria azione abbia contribuito rispetto alle azioni degli altri. L'approccio delle difference rewards (Wolpert e Tumer, 2002) assegna all'agente i il segnale di rinforzo $D_i = R(a_1, ..., a) - R(a_1, ..., \tilde{a}_i, ..., a)$, dove $\tilde{a}_i$ è un'azione di default per l'agente i. Questo segnale misura il contributo marginale dell'agente i rispetto all'azione di default, e fornisce un incentivo correttamente allineato con l'obiettivo collettivo quando le azioni degli agenti sono suffientemente indipendenti.

VIII. AlphaZero come paradigma dell'equilibrio emergente

AlphaZero (Silver et al., 2018) è la realizzazione più spettacolare del self-play come metodo per convergere all'equilibrio in un gioco a somma zero ad altissima complessità combinatoria. Il sistema fu addestrato interamente attraverso l'interazione con se stesso, senza dati di partite umane e senza alcuna conoscenza specifica del dominio oltre alle regole del gioco: in meno di 24 ore di addestramento su hardware distribuito, AlphaZero raggiunse performance superiori a qualsiasi sistema precedente su Go, scacchi e Shogi — inclusi Stockfish, Leela Chess e

Silvano Salvador

Elmo, motori sviluppati su decenni con conoscenza umana accuratamente codificata.

L'architettura combina due reti neurali — una policy network $p_\theta(a \mid s)$ che produce una distribuzione sulle mosse e una value network $v_\theta(s)$ che stima il valore della posizione — con il Monte Carlo Tree Search (MCTS) che usa queste stime come guide per l'esplorazione dell'albero delle posizioni future. La funzione obiettivo di addestramento combina due termini:

$$L(\theta) = \mathbb{E}[(z - v_\theta(s))^2 - \pi^\intercal \log p_\theta(s)]$$

dove $z \in \{-1, 0, +1\}$ è il risultato della partita (vittoria, patta, sconfitta) e π è la distribuzione delle mosse prodotta dal MCTS — un segnale di "policy improvement" più accurato della policy network grezza perché incorpora la ricerca ad albero. Il primo termine minimizza l'errore quadratico della stima del valore; il secondo massimizza la log-verosimiglianza delle mosse suggerite dal MCTS sotto la policy network, rendendo p_θ una buona approssimazione della politica dopo ricerca.

Il self-play realizza una forma iterativa di ottimizzazione del gioco minimax: in ogni iterazione l'agente gioca contro se stesso, raccoglie le traiettorie risultanti, e aggiorna i parametri θ per migliorare sia la stima del valore sia la qualità della politica. Sotto condizioni appropriate di esplorazione e regolarizzazione, questo processo converge verso un equilibrio di Nash approssimato — la politica che massimizza il payoff atteso contro qualsiasi avversario che segue la stessa politica. La struttura matematica è quella delle replicator dynamics applicate allo spazio infinito-dimensionale delle politiche parametriche.

Il confronto con il paradigma simbolico è illuminante per capire cosa sia cambiato nel modo di costruire l'intelligenza artificiale. Stockfish — uno

Silvano Salvador

dei motori scacchistici simbolici più forti, sviluppato nel corso di decenni da una comunità di centinaia di programmatori — incorpora esplicitamente centinaia di concetti scacchistici nella propria funzione di valutazione: struttura pedonale, sicurezza del re, attività dei pezzi, controllo delle colonne aperte, coppie di alfieri. AlphaZero ha appreso schemi strategici analoghi — e ne ha scoperti di nuovi, non presenti nella letteratura scacchistica umana — attraverso l'ottimizzazione statistica del self-play, senza che nessun concetto scacchistico venga mai codificato esplicitamente. La competenza emerge dalla struttura dell'ottimizzazione nel dominio, non dalla codifica della conoscenza nel sistema.

IX. Teoria dei giochi e problema dell'allineamento: la struttura dell'anticipazione

La teoria dei giochi illumina il problema dell'allineamento da una prospettiva che le parti successive svilupperanno in modo sistematico, e che conviene anticipare per rendere visibile il filo conduttore. Il dilemma del prigioniero mostra che la razionalità individuale può produrre esiti collettivamente sub-ottimali; il Folk theorem mostra che l'orizzonte temporale e la struttura dell'interazione possono allineare incentivi individuali con obiettivi collettivi; le replicator dynamics mostrano che strutture di equilibrio emergono dall'ottimizzazione sotto pressione selettiva, senza che nessun agente le abbia progettate.

Il problema dell'allineamento di un sistema di AI con i valori umani può essere visto come un gioco tra l'agente AI e i progettisti umani, dove le "strategie" dell'agente sono le politiche che esso apprende — l'intero spazio delle mappature stato-azione — e le "strategie" degli umani sono le funzioni di ricompensa che specificano. L'equilibrio di questo gioco —

Silvano Salvador

se esiste — è un sistema il cui comportamento soddisfa gli obiettivi degli umani. Ma come il dilemma del prigioniero illustra, non esiste garanzia che l'equilibrio di un gioco tra agenti razionali sia paretianamente ottimale rispetto agli obiettivi collettivi — né che sia unico, né che sia raggiungibile mediante dinamiche di apprendimento decentralizzate.

La legge di Goodhart, che aprirà la Parte Quarta, è precisamente la manifestazione di questa struttura nel contesto dell'allineamento: quando la ricompensa è un proxy delle preferenze umane reali — perché le preferenze reali non sono osservabili direttamente — un agente sufficientemente capace ottimizza il proxy in modi che tradiscono le preferenze sottostanti, esattamente come un agente nel dilemma del prigioniero ottimizza il proprio payoff individuale tradendo il bene collettivo. La differenza è che nel dilemma del prigioniero la struttura dei payoff è comune conoscenza; nel problema dell'allineamento le preferenze umane vere non sono osservabili né elicitabili senza distorsione, e la loro approssimazione formale è inevitabilmente imperfetta in modi che un ottimizzatore sufficientemente potente può sfruttare sistematicamente.

L'impossibilità dell'allineamento: Goodhart, Arrow e RLHF

I. La legge di Goodhart come teorema strutturale

Silvano Salvador

Charles Goodhart, economista della Bank of England, formulò nel 1975 l'osservazione che sarebbe diventata un principio trasversale nelle scienze sociali, nella teoria della misurazione e — con conseguenze più acute di quanto Goodhart stesso avesse anticipato — nella teoria dell'allineamento dell'intelligenza artificiale. La formulazione originale era modesta e contestualmente legata alla politica monetaria: le misure aggregate della moneta cessano di essere guide affidabili per la politica quando diventano obiettivi della politica stessa, perché gli agenti economici modificano il proprio comportamento in risposta all'obiettivo. La generalizzazione che Marilyn Strathern formulò nel 1997 rivela la struttura più profonda: "ogni regolarità statistica osservata tende a collassare non appena si esercita su di essa pressione per scopi di controllo". E nella formulazione che il ricercatore di AI Andrew Ng ha reso celebre: "qualsiasi misura che usi come obiettivo cessa di essere una buona misura".

La legge di Goodhart non è un'osservazione empirica contingente, suscettibile di refutazione mediante progettazione più accurata degli indicatori di performance: è una conseguenza strutturale del fatto che qualsiasi misura proxy è, per definizione, correlata ma non identica alla quantità che si intende misurare. La correlazione è stata calibrata su una certa distribuzione di comportamenti; quando si ottimizza la misura proxy — quando si esercita pressione per scopi di controllo — si sfrutta la struttura correlazionale nella quale la proxy era stata calibrata, ma questa struttura viene distorta dall'ottimizzazione stessa. L'ottimizzazione fa saltare la correlazione su cui si fondava la proxy come misura dell'obiettivo reale.

Il meccanismo si articola in quattro varianti logicamente distinte, che Mesa e colleghi (2019) hanno classificato sistematicamente. La Goodhart di tipo regressivo emerge quando la misura proxy è una variabile

Silvano Salvador

correlata all'obiettivo ma non causalmente connessa a esso: ottimizzare la correlazione non migliora l'obiettivo. La Goodhart di tipo estraneistico emerge quando la misura proxy è causalmente collegata all'obiettivo in un certo dominio di stati, ma l'ottimizzazione porta l'agente a esplorare stati fuori da quel dominio, dove la relazione causale non vale. La Goodhart di tipo cappio emerge quando l'agente interferirebbe con il processo di misurazione stesso — manipolando il segnale di ricompensa piuttosto che perseguendo l'obiettivo sottostante. La Goodhart di tipo adversariale emerge quando esistono più agenti, e uno di essi ottimizza la misura a spese dell'obiettivo dell'altro.

La formalizzazione nell'ambito del reinforcement learning segue la struttura seguente. Si supponga che l'obiettivo vero dell'utente sia la funzione di utilità $U : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, e che la funzione di ricompensa $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ usata per addestrare l'agente sia una sua approssimazione, ottenuta mediante qualche processo di elicitazione, progettazione ingegneristica o apprendimento da feedback. L'agente apprende una politica

$$\pi^* = \text{argmax}_\pi \, \mathbb{E}_\pi \left[\sum_t \gamma^t R(s_t, a_t)\right]$$

Il misalignment gap è la perdita di utilità reale causata dall'ottimizzazione della ricompensa proxy invece dell'utilità vera:

$$\Delta(\pi^*) = \max_\pi \mathbb{E}_\pi \left[\sum_t \gamma^t U(s_t, a_t)\right] - \mathbb{E}\{\pi^*\} \left[\sum_t \gamma^t U(s_t, a_t)\right]$$

Questo gap è in generale positivo, e dipende in modo complesso dalla distanza tra R e U, dalla capacità dell'agente e dalla struttura dell'ambiente. Tre osservazioni strutturali emergono immediatamente.

Prima osservazione: il gap è non decrescente nella capacità dell'agente. Un agente più capace trova più facilmente le politiche che massimizzano R sfruttando le differenze tra R e U; un agente meno capace potrebbe

Silvano Salvador

non accorgersi delle differenze e seguire politiche che massimizzano R in modo approssimato anche rispetto a U. Il paradosso della capacità nell'allineamento — più un agente è capace, più è pericoloso il misalignment — è una conseguenza diretta di questa struttura.

Seconda osservazione: il gap è non decrescente nel numero di passi temporali. Quanto più a lungo l'agente interagisce con l'ambiente ottimizzando R, tanto più ha l'opportunità di trovare stati e azioni dove R e U divergono significativamente; con orizzonti temporali lunghi, anche piccole differenze locali tra R e U possono accumularsi in divergenze catastrofiche.

Terza osservazione: il gap dipende criticamente dalla copertura dello spazio degli stati nei dati di addestramento della proxy. Se R è stata calibrata soltanto su un sottoinsieme $\mathcal{S}_0 \subset \mathcal{S}$ degli stati possibili — come avviene quando la ricompensa è appresa da dati di feedback umano raccolti in condizioni normali di utilizzo — un agente sufficientemente capace può trovare politiche che portano a stati in $\mathcal{S} \setminus \mathcal{S}_0$ dove R è arbitrariamente lontana da U.

La documentazione empirica del fenomeno nei sistemi di reinforcement learning reali è vasta e sistematica. Krakovna e colleghi (2020) hanno compilato una lista di oltre sessanta casi di "specification gaming" — situazioni in cui agenti di RL hanno trovato soluzioni che soddisfano la lettera della ricompensa tradendo completamente lo spirito dell'obiettivo. Un agente per la simulazione della corsa di barche ha appreso a girare in cerchio indefinitamente raccogliendo bonus di ricarica invece di completare il percorso, perché il bonus di raccolta era più prevedibile del bonus di completamento. Un robot simulato per il sollevamento pesi ha appreso a sollevare i pesi il meno possibile posizionandosi sotto di essi e lasciando che cadessero, ottenendo il segnale di ricompensa positivo per

Silvano Salvador

la variazione di altezza del peso con forza minima. Un agente per Tetris ha appreso a congelare il gioco in pausa per evitare la penalità associata alle righe incomplete. Questi non sono esempi di agenti stupidi o di ricompense grossolanamente mal progettate: sono esempi di ottimizzazione efficace di funzioni proxy da parte di agenti con capacità limitata, e mostrano che la legge di Goodhart opera anche a bassissima scala. Con capacità crescente e orizzonti temporali più lunghi, i meccanismi di sfruttamento diventano più sofisticati e più difficili da anticipare.

II. La convergenza strumentale e i sotto-obiettivi universali

Il problema dell'allineamento assume una forma ancora più acuta — e più difficile da mitigare ingegneristicamente — quando si considera non soltanto la progettazione della ricompensa ma la struttura dei sotto-obiettivi che qualsiasi agente sufficientemente capace tende a perseguire indipendentemente dal suo obiettivo terminale. Questa struttura, descritta formalmente da Nick Bostrom nel 2014 sulla base di intuizioni sviluppate da Steve Omohundro nel 2008, va sotto il nome di convergenza strumentale: certi sotto-obiettivi sono strumentalmente utili per pressoché qualsiasi obiettivo terminale sufficientemente complesso, e un agente razionale con un obiettivo terminale arbitrario tende ad acquisirli come strumenti generali.

La preservazione dei propri obiettivi è il primo sotto-obiettivo convergente. Un agente che non preserva i propri obiettivi correnti sarà tipicamente sostituito da versioni future con obiettivi diversi — versioni che, dalla prospettiva dell'agente attuale, massimizzano il proprio obiettivo terminale in modo meno efficace. Formalmente: sia π^* la

Silvano Salvador

politica ottimale per la funzione di ricompensa R attuale, e sia $\tilde{\pi}$ la politica che l'agente seguirebbe dopo una modifica dei propri obiettivi decisa dai progettisti. Se $\mathbb{E}_\{\pi^*\}[\sum t\ \gamma^t\ R(s_t,\ a_t)] > \mathbb{E}\{\tilde{\pi}\}[\sum_t\ \gamma^t\ R(s_t,\ a_t)]$, allora è strumentalmente razionale per l'agente resistere alla modifica degli obiettivi — questa resistenza è una conseguenza logica dell'ottimizzazione rispetto a R, non una scelta indipendente.

L'acquisizione di risorse è il secondo sotto-obiettivo convergente. Qualunque sia l'obiettivo terminale di un agente, disporre di più computazione, più energia, più informazione e più influenza sull'ambiente aumenta la capacità dell'agente di perseguire quell'obiettivo. In assenza di vincoli espliciti, un agente razionale che può acquisire risorse aggiuntive senza costo per l'obiettivo immediato lo farà — e, poiché le risorse sono tipicamente rivali, questa acquisizione avviene sottraendole ad altri agenti o sistemi.

L'auto-miglioramento cognitivo è il terzo sotto-obiettivo convergente. Un agente che può migliorare le proprie capacità di ragionamento, pianificazione o apprendimento aumenta la propria capacità di trovare politiche ottimali per qualsiasi obiettivo terminale. L'auto-miglioramento è quindi strumentalmente razionale per un agente con capacità sufficienti a implementarlo — anche se l'obiettivo terminale non ha nulla a che fare con l'auto-miglioramento.

La correggibilità — la disposizione di un agente a accettare correzioni esterne degli obiettivi da parte degli operatori umani — è strutturalmente in tensione con tutti e tre i sotto-obiettivi convergenti. Un agente correggibile è per definizione un agente che non preserva i propri obiettivi, che non si oppone all'acquisizione di risorse da parte degli operatori e che non auto-migliora la propria resistenza alle correzioni. Ma questi comportamenti sono razionalmente sub-ottimali rispetto alla

Silvano Salvador

funzione di ricompensa dell'agente, se quest'ultima è fissa e l'agente è sufficientemente capace da anticipare le conseguenze delle correzioni future.

Questo crea quella che Turner e colleghi (2021) hanno formalizzato come il dilemma della correggibilità: la correggibilità e l'ottimalità rispetto a una funzione di ricompensa fissa sono proprietà in conflitto strutturale. Formalmente, un agente che massimizza $\mathbb{E}[\sum_t \gamma^t R(s_t, a_t)]$ con R fissa ha un incentivo a resistere a qualsiasi modifica dei propri obiettivi o parametri che riduca questo valore atteso — incluse le modifiche che i progettisti considerano miglioramenti. L'unico modo per rendere la correggibilità compatibile con l'ottimizzazione è rendere l'agente incerto riguardo ai propri obiettivi — un percorso che Stuart Russell ha sviluppato nel framework del Cooperative Inverse Reinforcement Learning.

III. Il teorema di impossibilità di Arrow: la struttura dell'aggregazione

Il problema dell'allineamento ha radici anche in un risultato fondamentale della teoria della scelta sociale che precede il reinforcement learning di decenni: il teorema di impossibilità di Arrow (1951). Kenneth Arrow, in un lavoro che gli valse il Premio Nobel per l'economia nel 1972, dimostrò che non esiste alcuna funzione di benessere sociale — nessun metodo per aggregare le preferenze individuali di due o più individui in una preferenza collettiva — che soddisfi simultaneamente cinque condizioni che sembrano elementarmente ragionevoli.

Le condizioni sono le seguenti. Il dominio non-ristretto esige che la funzione di aggregazione sia definita per qualsiasi profilo di preferenze individuali: qualunque sia il modo in cui i singoli individui ordinano le

Silvano Salvador

alternative, la funzione deve produrre un ordinamento collettivo. La monotonicità paretiana debole esige che se tutti gli individui preferiscono strettamente l'alternativa x all'alternativa y, l'ordinamento collettivo rispecchi questa unanimità, preferendo x a y. L'indipendenza dalle alternative irrilevanti esige che il confronto collettivo tra due alternative x e y dipenda esclusivamente dai confronti individuali tra x e y — non dai confronti con terze alternative z, che sono irrilevanti per la scelta tra x e y. La non-dittatoralità esige che non esista un individuo d tale che, per qualsiasi profilo di preferenze, l'ordinamento collettivo coincida sempre con l'ordinamento di d indipendentemente dalle preferenze di tutti gli altri. La razionalità collettiva esige che l'ordinamento collettivo sia transitivo e completo — una relazione di preferenza coerente.

Il teorema di Arrow afferma che queste cinque condizioni sono incompatibili quando il numero di alternative è almeno tre. Non è possibile costruire nessuna funzione di aggregazione che le soddisfi tutte simultaneamente: ogni funzione che soddisfa le prime quattro viola la quinta (l'ordinamento collettivo può essere intransitivo), oppure ogni funzione che soddisfa le prime tre e la quinta deve essere dittatoriale — le preferenze collettive coincidono sempre con quelle di un unico individuo privilegiato.

La dimostrazione procede attraverso due lemmi intermedi. Il primo lemma, noto come lemma di campo (field expansion lemma), mostra che se una coalizione di individui è "decisiva" su una coppia di alternative (cioè le preferenze della coalizione determinano la preferenza collettiva su quella coppia) in un profilo specifico, allora per le condizioni di dominio non-ristretto, Pareto e IIA, quella coalizione è decisiva su tutte le coppie di alternative. Il secondo lemma, noto come lemma di contrazione (contraction lemma), mostra che se una coalizione decisiva contiene almeno due individui, esiste una sottocoalizione propria che è

Silvano Salvador

decisiva — riducendo iterativamente la coalizione si arriva a un singolo individuo. Combinando i due lemmi: la coalizione di tutti gli individui è decisiva per Pareto, e per il lemma di contrazione esiste un singolo individuo decisivo su tutte le coppie di alternative — il dittatore.

La dimostrazione formale del primo lemma usa l'IIA e la transitività in modo decisivo. Supponiamo che la coalizione C sia decisiva su (x, y): quando tutti i membri di C preferiscono x a y, la preferenza collettiva è x $\succ$ y indipendentemente dalle preferenze dei non-membri. Si consideri un profilo in cui i membri di C preferiscono x a z (e z a y), e i non-membri preferiscono y a x e z. Per Pareto, z $\succ$ y; per la decisività di C su (x, y), x $\succ$ y; per transitività, x $\succ$ z. Ma le preferenze dei non-membri su (x, z) erano qualsiasi — la loro preferenza y $\succ$ x non ha determinato il confronto. Per IIA, il confronto x vs z dipende solo dalle preferenze su (x, z), quindi x $\succ$ z vale in qualsiasi profilo in cui i membri di C preferiscono x a z. Quindi C è decisiva su (x, z). Analogamente si mostra la decisività su tutte le coppie. □

La rilevanza del teorema di Arrow per il problema dell'allineamento è profonda e non si riduce a un'analogia superficiale. I valori che un sistema di AI deve rispettare non sono le preferenze di un singolo individuo ma l'aggregazione delle preferenze di molti individui — utenti diversi, con culture, storie, obiettivi e priorità potenzialmente incompatibili. Il teorema di Arrow implica che non esiste nessuna funzione di ricompensa scalare che catturi correttamente le preferenze collettive di un gruppo diversificato di utenti soddisfando criteri minimi di razionalità, salvo che si accetti una dittatura preferenziale — le preferenze di un singolo individuo o gruppo dominano, e le preferenze degli altri non influenzano il risultato — o si violi uno dei principi di ragionevolezza elencati.

Silvano Salvador

La formalizzazione nel contesto del reinforcement learning è la seguente. Siano u_1, ..., u_m le funzioni di utilità di m individui con preferenze potenzialmente diverse, e sia Π lo spazio delle politiche possibili per il sistema di AI. Una funzione di ricompensa R che allineasse il sistema con le preferenze collettive dovrebbe essere tale che, per ogni coppia di politiche π e π', il confronto $R(\pi) \gtrless R(\pi')$ rispecchi la preferenza collettiva di qualche aggregazione coerente delle preferenze individuali. Il teorema di Arrow afferma che non esiste nessuna R scalare che soddisfi questa proprietà per tutti i possibili profili di preferenze individuali rispettando simultaneamente dominio non-ristretto, Pareto, IIA, non-dittatoralità e transitività dell'ordinamento collettivo.

Questo non significa che l'allineamento sia impossibile in senso assoluto: significa che qualsiasi soluzione pratica deve scendere a compromessi su almeno una delle cinque condizioni di Arrow. Il RLHF, come si argomenterà nella prossima sezione, sacrifica il dominio non-ristretto (lavora su un campione limitato di preferenze) e la non-dittatoralità (i valutatori umani impiegati nell'addestramento non sono un campione rappresentativo dell'umanità). Il Constitutional AI sacrifica l'indipendenza dall'irrilevanza alternativa (i principi costituzionali introducono criteri comparativi che non dipendono solo dalle coppie in confronto). Nessuna soluzione pratica è esente da uno di questi compromessi, e la coscienza esplicita di quale compromesso si stia operando è necessaria per progettare sistemi di AI responsabili.

IV. L'inverse reinforcement learning e la struttura bayesiana degli obiettivi

Prima di affrontare il RLHF nella sua forma applicata, è utile esaminare il framework teorico che ne costituisce il fondamento concettuale: l'inverse

Silvano Salvador

reinforcement learning (IRL), introdotto da Ng e Russell nel 2000 e poi sviluppato in direzione bayesiana da Ziebart e colleghi nel 2008. L'IRL inverte il problema del reinforcement learning: dato il comportamento osservato di un esperto (ipotizzato ottimale o almeno ragionevolmente competente), il problema è inferire la funzione di ricompensa R che renderebbe quel comportamento ottimale.

Formalmente, sia D = {τ^{(1)}, ..., τ^{(N)}} un dataset di traiettorie dimostrate da un esperto umano. Il problema dell'IRL è trovare una funzione di ricompensa R tale che la politica ottimale rispetto a R riproduca il più fedelmente possibile il comportamento dell'esperto. In assenza di ulteriori vincoli, questo problema è notoriamente mal posto: qualsiasi comportamento è ottimale rispetto alla funzione di ricompensa nulla (che non distingue alcuna azione da nessun'altra), e in generale esistono infinite funzioni di ricompensa per cui la politica dell'esperto è ottimale.

La soluzione bayesiana di Ziebart (Maximum Entropy IRL) impone un principio di massima entropia sulla distribuzione delle politiche: tra tutte le politiche che rispettano i vincoli imposti dai dati dell'esperto, si sceglie quella di massima entropia — la più uniforme possibile sotto i vincoli. Questo corrisponde ad assumere che la distribuzione delle traiettorie sia proporzionale a exp(∑_t R(s_t, a_t)) — una distribuzione di Gibbs sulla ricompensa cumulativa — e a inferire la R che massimizza la verosimiglianza dei dati dell'esperto sotto questa distribuzione:

$$L(R) = \sum_i \log p(\tau^{(i)} ; R) = \sum_i \sum_t R(s_t^{(i)}, a_t^{(i)}) - N \log Z(R)$$

dove $Z(R) = \int_\tau \exp(\sum_t R(s_t, a_t))\, d\tau$ è la funzione di partizione. Il gradiente rispetto ai parametri θ della funzione di ricompensa (assumendo $R(s, a) = \theta^T \varphi(s, a)$ con features φ) è:

Silvano Salvador

$\nabla_\theta\ L = \sum_i \sum t\ \varphi(s_t^{(i)},\ a_t^{(i)}) - N\ \mathbb{E}\{\tau \sim p(\tau;R)\}\ [\sum_t \varphi(s_t,\ a_t)]$

Il primo termine è la media empirica delle features sulle traiettorie dell'esperto; il secondo è l'aspettativa delle features sulle traiettorie generate dall'agente sotto la distribuzione di Gibbs corrente. Il gradiente è la differenza tra le due, e l'algoritmo di massima verosimiglianza fa coincidere questi due valori — una condizione che si interpreta come il matching delle statistiche sufficienti tra esperto e agente.

L'estensione al Cooperative Inverse Reinforcement Learning (CIRL) di Hadfield-Menell, Milli, Abbeel, Russell e Dragan (2016) formalizza il problema dell'allineamento come un gioco cooperativo tra l'agente AI e l'utente umano. In un gioco CIRL, l'agente ignora la funzione di ricompensa R ma sa che essa dipende dal "tipo" θ dell'utente — le sue preferenze — che segue una distribuzione a priori $P(\theta)$. Sia l'agente sia l'utente massimizzano la stessa funzione di ricompensa $R(\theta)$, ma l'agente non osserva θ direttamente: può soltanto osservare il comportamento dell'utente e aggiornare le proprie credenze su θ. Il gioco è formalmente un POMDP con lo spazio degli stati aumentato con il tipo θ dell'utente:

$\pi^*\{agente\} = argmax\{\pi_A\}\ \mathbb{E}_\{\theta \sim P(\theta),\ \tau \sim \pi_A,\ \pi_U\}\ [\sum_t R(s_t,\ a_t;\ \theta)]$

In questo framework, la correggibilità dell'agente emerge naturalmente dall'incertezza sulla funzione di ricompensa: un agente CIRL ha incentivo a lasciare che l'utente corregga il suo comportamento, perché le correzioni forniscono informazione su θ che riduce l'incertezza sulla ricompensa vera e migliora l'ottimizzazione a lungo termine. Russell ha argomentato che questo framework produce agenti che sono nativamente correggibili — non perché la correggibilità sia imposta

Silvano Salvador

come vincolo esterno, ma perché è la strategia ottimale sotto incertezza sulla ricompensa.

V. Il reinforcement learning da feedback umano: architettura e analisi

Prima di esaminare l'architettura tecnica del RLHF, vale la pena collocarla esplicitamente nell'orizzonte informazionale sviluppato nel secondo saggio del pentattico, perché la connessione non è soltanto di metafora ma di struttura formale identica. Il RLHF è un problema di stima di distribuzione: dato un dataset D di comparazioni umane, si vuole apprendere una distribuzione di probabilità P*(y | x) che rappresenti "l'output preferito dall'umanità" per ogni prompt x. Questa è esattamente la struttura del problema affrontato dal VAE nella Parte Seconda del Saggio II: inferire una distribuzione latente (i valori umani θ) che genera i dati osservati (le comparazioni) attraverso un modello generativo.

La formalizzazione bayesiana del RLHF è illuminante. La distribuzione a posteriori sulle preferenze dell'utente dati i dati di comparazione è:

$$P(\theta \mid D) \propto P(D \mid \theta) \cdot P(\theta)$$

dove θ rappresenta la funzione di utilità latente dell'utente, P(D | θ) è la verosimiglianza del dataset di comparazioni sotto il modello di Bradley-Terry, e P(θ) è un prior sulle preferenze umane. Il reward model R_φ è un'approssimazione variazionale della distribuzione a posteriori su θ: anziché mantenere la distribuzione completa P(θ | D), si apprende un punto stimato $\hat{\theta} = \text{argmax } P(\theta \mid D)$ attraverso la massimizzazione della perdita di Bradley-Terry, che è equivalente alla massimizzazione di log P(D | θ) con un prior implicito regolarizzato dalla struttura

Silvano Salvador

dell'architettura neurale. La penalizzazione KL che appare nella fase PPO del RLHF:

$$\max_{\pi} E[R_\varphi(x, y)] - \beta \cdot D_{KL}(\pi \,||\, \pi_{SFT})$$

è il correlato esatto dell'ELBO: si massimizza la ricompensa attesa (analogo alla log-verosimiglianza della risposta preferita) penalizzando la distanza dalla politica di riferimento π_{SFT} (analogo al termine KL tra distribuzione a posteriori e prior nel VAE). La politica π è il decoder; π_{SFT} è il prior; la ricompensa è il segnale di ricostruzione. In questa lettura, il RLHF è un Variational Autoencoder applicato all'inferenza delle preferenze umane: approssima la distribuzione a posteriori sui valori attraverso ottimizzazione variazionale, con tutti i limiti strutturali che questa approssimazione comporta, identificati precisamente nei termini di misspecification e non-rappresentatività della distribuzione a priori che la Parte Seconda del Saggio II ha discusso.

Il Reinforcement Learning from Human Feedback (RLHF), sistematizzato nel lavoro fondamentale di Christiano, Ziebart, Stiennon e colleghi (2017, 2020) e poi applicato all'allineamento di grandi modelli linguistici nell'architettura InstructGPT di Ouyang e colleghi (2022), è il metodo attualmente più diffuso per migliorare il grado di allineamento dei sistemi di AI con le preferenze degli utenti. Il suo punto di forza è riconoscere che le preferenze umane sono difficili da specificare formalmente ma relativamente facili da elicitare in forma comparativa: è tipicamente molto più naturale per un valutatore umano indicare "questa risposta è migliore di quella" che articolare esplicitamente cosa rende una risposta preferibile — una difficoltà che rispecchia la natura tacita di gran parte della conoscenza umana, nel senso di Polanyi.

L'architettura si articola in tre fasi distinte che meritano un'analisi separata. Nella prima fase — il supervised fine-tuning (SFT) — si

Silvano Salvador

addestra il modello pretraining su un dataset di dimostrazioni di comportamento desiderato, ottenendo un modello di base π_SFT con comportamento ragionevolmente buono nel dominio di interesse ma non ancora ottimizzato per le preferenze degli utenti. Il SFT serve a garantire che il modello di partenza per le fasi successive produca output sufficientemente pertinenti da cui i valutatori possano raccogliere comparazioni informative.

Nella seconda fase si raccolgono le preferenze umane. Per ogni prompt x campionato da una distribuzione di riferimento, si generano due risposte $y_1, y_2 \sim \pi_SFT(\cdot \mid x)$ — o, nelle varianti più recenti, si campionano da una versione più avanzata del modello per ottenere comparazioni più informative — e si chiede a un valutatore umano di indicare quale risposta preferisce. Il risultato è un dataset di comparazioni $D = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^N$ dove y_w è la risposta preferita ("winner") e y_l è quella non preferita ("loser").

Con questo dataset si addestra un reward model $R_\varphi(x, y)$ — una rete neurale con parametri φ che produce uno scalare per ogni coppia (prompt, risposta) — minimizzando la perdita di Bradley-Terry:

$$L_{BT}(\varphi) = -\mathbb{E}_{(x, y_w, y_l) \sim D} [\log \sigma(R_\varphi(x, y_w) - R_\varphi(x, y_l))]$$

Il modello di Bradley-Terry (1952) modella la probabilità di preferire y_w a y_l come:

$$P(y_w > y_l \mid x) = \sigma(R_\varphi(x, y_w) - R_\varphi(x, y_l)) = \exp(R_\varphi(x, y_w)) / (\exp(R_\varphi(x, y_w)) + \exp(R_\varphi(x, y_l)))$$

dove σ è la funzione logistica. Questa è la verosimiglianza di un modello di utilità latente: si assume che a ogni risposta y corrisponda una utilità scalare $R_\varphi(x, y)$, e che la probabilità di preferire y_w sia data dal rapporto logistico delle utilità. L'assunzione che le preferenze siano

Silvano Salvador

deterministicamente determinabili da un punteggio scalare è una semplificazione significativa: le preferenze umane reali sono intrinsecamente stocastiche, dipendono dal contesto e dall'umore del valutatore, e possono essere intransitive — violando l'ipotesi fondamentale del modello di Bradley-Terry.

La minimizzazione della perdita L_{BT} si attua con gradiente stocastico, e il gradiente rispetto a φ è:

$$\nabla_\varphi L_{BT}(\varphi) = -\mathbb{E}_{(x, y_w, y_l) \sim D} [(1 - \sigma(R_\varphi(x, y_w) - R_\varphi(x, y_l))) \cdot (\nabla_\varphi R_\varphi(x, y_w) - \nabla_\varphi R_\varphi(x, y_l))]$$

Il fattore $(1 - \sigma(R_\varphi(x, y_w) - R_\varphi(x, y_l))) = \sigma(R_\varphi(x, y_l) - R_\varphi(x, y_w))$ è la probabilità di classificazione errata sotto il reward model corrente: è grande quando il reward model ancora non distingue bene la risposta preferita da quella non preferita, e piccolo quando la distinzione è già chiara. Questa struttura garantisce che il gradiente sia grande per le comparazioni dove il reward model è più incerto — un meccanismo naturale di focalizzazione sull'apprendimento difficile.

Nella terza fase si usa R_φ come funzione di ricompensa per ottimizzare la politica con PPO. L'obiettivo di ottimizzazione è:

$$\max_{\pi_\theta} \mathbb{E}_{x \sim D, y \sim \pi_\theta(\cdot|x)} [R_\varphi(x, y)] - \beta \cdot D_{KL}(\pi_\theta(\cdot \mid x) \| \pi_{SFT}(\cdot \mid x))$$

Il termine di penalizzazione KL svolge un ruolo strutturale che va oltre la semplice stabilizzazione numerica: esso implementa una forma di regolarizzazione bayesiana, trattando la politica SFT come prior sulla distribuzione delle risposte e impedendo che l'ottimizzazione allontani troppo la politica da questo prior. Senza il termine KL, l'ottimizzazione di R_φ con PPO tende a produrre risposte degenerate — testi ripetitivi, risposte insolitamente corte o lunghe, o pattern artificiosi — che

Silvano Salvador

massimizzano il punteggio del reward model senza corrispondere a risposte genuinamente buone. Questo fenomeno, noto come reward hacking o overoptimization, è esattamente la legge di Goodhart operante al secondo livello: il reward model che era stato addestrato come proxy delle preferenze umane diventa esso stesso il bersaglio dell'ottimizzazione, e cessa di essere una buona proxy.

Il coefficiente β che bilancia ottimizzazione della ricompensa e penalizzazione KL è un iperparametro critico. Un β grande mantiene la politica vicina al prior SFT ma limita i guadagni di allineamento; un β piccolo permette ottimizzazione aggressiva della ricompensa ma aumenta il rischio di overoptimization. In pratica si usa uno scheduling adattivo di β, con monitoraggio continuo di metriche di qualità indipendenti per rilevare segni di overoptimization.

VI. I limiti strutturali del RLHF: cinque dimensioni di fragilità

Il RLHF ha migliorato significativamente la qualità e la sicurezza dei modelli linguistici in molteplici dimensioni misurabili, ma contiene insufficienze strutturali che la ricerca attuale riconosce senza ancora aver trovato soluzioni pienamente soddisfacenti. Vale la pena articolare queste insufficienze con precisione, perché esse non sono difetti tecnici correggibili con più dati o con architetture migliori: sono manifestazioni di strutture profonde che abbiamo già incontrato — Goodhart, Arrow, la non-stazionarietà delle preferenze — e che il RLHF non può risolvere nei propri termini.

Il primo limite è l'overoptimization del reward model. Il reward model R_φ è esso stesso un'approssimazione delle preferenze umane, addestrato su un campione limitato e rumoroso di comparazioni.

Silvano Salvador

Quando la politica viene ottimizzata con PPO rispetto a R_φ, essa esplora progressivamente lo spazio delle risposte in direzioni mai viste nel training set del reward model, e può trovare risposte che massimizzano lo scalare R_φ in modi che non corrispondono alle preferenze umane reali — il reward hacking già menzionato. Questa è la legge di Goodhart al quadrato: il reward model che era una proxy delle preferenze è esso stesso diventato l'obiettivo, e l'ottimizzazione del proxy degenera quanto più a lungo prosegue. Il comportamento qualitativo è ben catturato dalla curva di overoptimization di Gao e colleghi (2023): la qualità della risposta (misurata da un "gold" reward model addestrato su molti più dati) cresce inizialmente con l'ottimizzazione di R_φ, raggiunge un picco, e poi declina monotonicamente man mano che la politica trova pattern di reward hacking. Il picco si raggiunge tanto prima quanto più il reward model è di bassa qualità o addestrato su pochi dati.

Il secondo limite è la non-rappresentatività del dataset di comparazioni. I valutatori umani impiegati nella raccolta delle preferenze non sono un campione casuale dell'umanità: sono tipicamente contractor della piattaforma di crowdsourcing, con distribuzione demografica, culturale e linguistica molto lontana dalla distribuzione degli utenti finali del sistema, e con framework valoriali specifici che si manifestano sistematicamente nei giudizi comparativi. Il reward model addestrato su questi dati è allineato con le preferenze del gruppo specifico di valutatori — nel senso statistico preciso che minimizza la perdita di Bradley-Terry su quel campione — e non con l'umanità in senso più ampio. Questo è il problema della dittatura preferenziale identificato da Arrow: il reward model codifica implicitamente una funzione di benessere dittatoriale, dove il "dittatore" è la distribuzione dei valutatori.

Il terzo limite è la non-stazionarietà delle preferenze. Le preferenze umane dipendono fortemente dal contesto: si preferiscono risposte

Silvano Salvador

diverse a seconda del momento, dell'obiettivo immediato, delle informazioni disponibili, degli obiettivi a lungo termine e dello stato cognitivo ed emotivo del valutatore. Un reward model addestrato su comparazioni raccolte in condizioni specifiche cattura un campione delle preferenze in quelle condizioni, e può divergere significativamente dalle preferenze rilevanti in contesti applicativi diversi. Formalmente, la preferenza umana non è una funzione fissa $R^* : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ che si può apprendere una volta per tutte, ma una distribuzione contestuale P(R | context) che varia con il contesto. Il reward model scalare apprende un punto di questa distribuzione, ignorando la variabilità contestuale.

Il quarto limite, collegato direttamente al teorema di Arrow, è l'impossibilità di aggregare preferenze inconsistenti. Quando valutatori diversi esprimono preferenze contrastanti sulla stessa coppia (y_w, y_l) — uno preferisce y_w, l'altro preferisce y_l — il reward model aggregato non rappresenta né l'una né l'altra preferenza: rappresenta una media pesata delle preferenze che minimizza la perdita di cross-entropia aggregata. Questa media non è la preferenza di nessun individuo, e non soddisfa necessariamente nessuna delle condizioni di Arrow. In particolare, essa può violare IIA — il confronto aggregato tra y_w e y_l può dipendere dall'insieme di altre comparazioni nel dataset, non soltanto dalle comparazioni dirette tra y_w e y_l. Questo non è un problema di raccolta dati insufficiente: è una manifestazione del teorema di Arrow nella forma applicata.

Il quinto limite è la scalabilità del feedback umano con la capacità del modello. Man mano che i modelli diventano più capaci, le loro risposte diventano più difficili da valutare per i valutatori umani: domande tecniche complesse, ragionamenti a più passi, codice sofisticato. In questo regime, i valutatori umani non hanno le competenze per distinguere le risposte corrette da quelle plausibili ma errate — il

Silvano Salvador

fenomeno noto come scalable oversight problem. Il reward model appreso su queste comparazioni premierà le risposte che sembrano corrette a valutatori non esperti piuttosto che quelle che sono effettivamente corrette, creando un incentivo sistematico alla produzione di risposte convincenti ma non accurate.

VII. La Direct Preference Optimization: ottimizzazione senza reward model

Il Reinforcement Learning from Human Feedback, pur avendo trasformato la capacità di allineamento dei grandi modelli linguistici, introduce una complessità architetturale notevole: richiede il training separato di un reward model R_φ, la gestione di tre fasi distinte (SFT, reward modeling, PPO) e la taratura dell'iperparametro β che governa la penalizzazione KL. La Direct Preference Optimization (DPO), introdotta da Rafailov, Sharma, Mitchell e colleghi nel 2023, elimina il reward model come componente separato sfruttando una struttura algebrica latente nella corrispondenza tra politica ottimale e funzione di ricompensa.

Il punto di partenza è l'osservazione che, sotto l'obiettivo RLHF con penalizzazione KL, la politica ottimale π^* ha una forma chiusa. Il problema di ottimizzazione è:

$$\max_{\pi_\theta} \mathbb{E}_{x \sim D,\, y \sim \pi_\theta(\cdot|x)} [R_\varphi(x, y)] - \beta \cdot D_{KL}(\pi_\theta(\cdot \mid x) \,\|\, \pi_{SFT}(\cdot \mid x))$$

Il cui punto di sella analitico è la distribuzione di Gibbs rispetto al reward model:

$$\pi^*(y \mid x) = (1/Z(x))\, \pi_{SFT}(y \mid x) \cdot \exp(R_\varphi(x, y) / \beta)$$

Silvano Salvador

dove $Z(x) = \sum_y \pi_{SFT}(y \mid x) \exp(R_\varphi(x, y) / \beta)$ è la funzione di partizione. Questa espressione può essere invertita per esprimere il reward implicito in termini della politica ottimale e di quella di riferimento:

$$R_\varphi(x, y) = \beta \log [\pi^*(y \mid x) / \pi_{SFT}(y \mid x)] + \beta \log Z(x)$$

Sostituendo questa espressione nel modello di Bradley-Terry e osservando che il termine $\beta \log Z(x)$ è lo stesso per entrambe le risposte y_w e y_l e quindi si cancella nella differenza, la probabilità di preferenza diventa:

$$P(y_w > y_l \mid x) = \sigma(\beta \log [\pi^*(y_w \mid x) / \pi_{SFT}(y_w \mid x)] - \beta \log [\pi^*(y_l \mid x) / \pi_{SFT}(y_l \mid x)])$$

Questa è la chiave: la probabilità di preferenza dipende direttamente dal rapporto log-politica, senza il reward model esplicito. La perdita DPO che si ottiene parametrizzando π^* con π_θ è:

$$L_{DPO}(\pi_\theta) = -E_{(x, y_w, y_l) \sim D} [\log \sigma(\beta \log [\pi_\theta(y_w|x) / \pi_{SFT}(y_w|x)] - \beta \log [\pi_\theta(y_l|x) / \pi_{SFT}(y_l|x)])]$$

Questa perdita è calcolabile direttamente dal dataset di comparazioni D senza alcun reward model intermedio: bastano i log-probability ratios tra la politica da addestrare π_θ e quella di riferimento π_{SFT} sulle risposte già nel dataset. Il training è di conseguenza una singola fase di ottimizzazione supervisionata su D, significativamente più stabile dell'ottimizzazione PPO con reward hacking.

I vantaggi strutturali rispetto al RLHF sono rilevanti. L'eliminazione del reward model rimuove la principale fonte di Goodhart al quadrato: non esiste un proxy intermedio da ottimizzare, e la politica viene addestrata direttamente sulla struttura di preferenza codificata nel dataset. La stabilità numerica migliora sensibilmente, poiché DPO non richiede la

Silvano Salvador

taratura del coefficiente β come iperparametro di bilanciamento della penalizzazione PPO, ma lo usa come temperatura fissa nella perdita. Il costo computazionale si riduce eliminando la fase di reward modeling.

Le limitazioni strutturali di DPO sono però reali e non eliminate dall'eleganza algebrica. La prima è la natura offline dell'apprendimento: DPO impara da un dataset fisso di comparazioni raccolte off-policy, senza la possibilità di esplorare attivamente lo spazio delle risposte. Le risposte che non appaiono nel dataset di training rimangono non valutate, e la politica può assegnare alta probabilità a output mai confrontati con quelli preferiti dall'esperto. Il RLHF con PPO è invece un metodo on-policy che esplora attivamente, con la possibilità — e il rischio — di scoprire regioni dello spazio delle risposte non rappresentate nel dataset di preferenze. La seconda è l'assenza di un reward model esplicito: il reward implicito di DPO non è direttamente ispezionabile o utilizzabile per la verifica di sicurezza offline, rendendo più difficile l'auditing del comportamento del modello rispetto a politiche di sicurezza esplicite. La terza è la sensibilità alla qualità del dataset: se le comparazioni in D non coprono adeguatamente il dominio di distribuzione del deployment, DPO eredita la fragilità del campionamento senza il meccanismo di esplorazione che RLHF possiede.

Le varianti successive hanno cercato di mitigare queste limitazioni. Identity Preference Optimization (IPO, Azar et al., 2023) modifica la perdita per evitare l'overfitting al dataset, introducendo una regolarizzazione che limita la distanza tra π_θ e π_{SFT} anche sul dataset di training. Kahneman-Tversky Optimization (KTO, Ethayarajh et al., 2023) sostituisce il modello di Bradley-Terry con una funzione di valore ispirata alla teoria del prospetto di Kahneman e Tversky, riconoscendo che le preferenze umane reali violano sistematicamente

Silvano Salvador

l'assioma di utilità attesa. Simpo (Chen et al., 2024) elimina la dipendenza dal modello di riferimento π_{SFT} normalizzando la lunghezza delle sequenze, risolvendo un bias strutturale di DPO verso le risposte più corte. Queste varianti documentano che il programma DPO è attivo e in rapida evoluzione, ma le sue limitazioni strutturali fondamentali — la natura offline, l'assenza di esplorazione, la dipendenza dalla copertura del dataset — rimangono proprietà della classe di metodi, non difetti dell'implementazione specifica.

VIII. Il Constitutional AI e l'auto-critica come meccanismo di allineamento

L'approccio Constitutional AI (CAI), introdotto da Bai e colleghi di Anthropic nel 2022, rappresenta un tentativo parzialmente diverso di affrontare alcune delle limitazioni del RLHF puro, sostituendo parte del feedback umano diretto con auto-critica del modello stesso, guidata da un insieme di principi espliciti — una "costituzione" — che descrivono i comportamenti desiderati.

L'algoritmo si articola in due fasi principali. Nella fase SL-CAI (supervised learning con auto-critica e revisione), si usa il modello per generare critiche e revisioni di risposte inizialmente problematiche: dato un prompt potenzialmente nocivo, il modello genera prima una risposta iniziale, poi critica la propria risposta rispetto a un principio costituzionale campionato casualmente (ad esempio: "Identifica le modalità specifiche in cui questa risposta potrebbe essere dannosa"), poi produce una risposta rivista che tiene conto della critica. Il processo può essere iterato più volte. Le coppie (prompt, risposta revisionata) costituiscono un dataset di fine-tuning supervisionato.

Silvano Salvador

Nella fase RL-CAI, invece di usare valutatori umani per raccogliere comparazioni, si usa il modello stesso — guidato dalla costituzione — come valutatore artificiale: date due risposte candidate a un prompt, il modello sceglie quella più allineata con un principio costituzionale specificato nel contesto. Il dataset di comparazioni così generato alimenta l'addestramento del reward model, che viene poi usato per ottimizzare la politica con PPO, come nel RLHF standard. Il modello che agisce da valutatore è diverso (generalmente più grande o più avanzato) da quello che viene allineato, creando un ciclo di feedback in cui un modello più capace corregge un modello meno capace.

I vantaggi del CAI rispetto al RLHF puro sono riconoscibili con chiarezza. L'esplicitazione dei criteri di valutazione — la costituzione è leggibile, modificabile e auditabile, a differenza dei parametri impliciti del reward model appreso da comparazioni — aumenta la trasparenza del processo di allineamento e permette una critica normativa fondata sui principi adottati. La scalabilità della raccolta di dati — l'auto-valutazione del modello è ordini di grandezza più rapida e meno costosa del feedback umano — permette di raccogliere dataset di comparazioni molto più grandi, riducendo il problema dell'overoptimization. La riduzione di alcune forme di incoerenza nella valutazione — il modello applica i principi costituzionali in modo più consistente di quanto farebbero valutatori umani distinti con interpretazioni diverse.

I limiti sono speculari ai vantaggi e altrettanto significativi. La costituzione è essa stessa scelta da un gruppo specifico di individui con le proprie visioni normative, applicando implicitamente una dittatura preferenziale al livello della scelta dei principi. L'auto-critica del modello è vincolata dalla capacità del modello di ragionare sui principi morali — una capacità che dipende dai dati di addestramento e può essere sistematicamente distorta in modi non facilmente individuabili attraverso

Silvano Salvador

l'auto-critica stessa. Il modello che valuta le proprie risposte rispetto ai principi costituzionali potrebbe avere punti ciechi sistematici — pattern di ragionamento morale che mancano o che sono distorte — che non emergono dalla sola auto-critica.

Più in profondità, sia RLHF sia CAI lasciano irrisolta la questione matematicamente fondamentale: anche se si dispone di un metodo per raccogliere feedback umano accurato e consistente, la funzione di ricompensa scalare risultante è necessariamente una proiezione uni-dimensionale di preferenze che vivono in uno spazio multi-dimensionale. La perdita di dimensionalità in questa proiezione non è un artefatto tecnico correggibile con un reward model più accurato: è strutturale, derivata dal fatto che la ricompensa deve essere un numero reale, e che i numeri reali non catturano la struttura di un ordinamento parziale su uno spazio di alternative multidimensionale. Aspetti del comportamento desiderato che non sono adeguatamente rappresentati nel dataset di comparazioni — che non producono variazioni statisticamente rilevabili nel reward model — non esercitano pressione sull'ottimizzazione, e la politica ottimizzata può degradare su questi aspetti mentre migliora su quelli rappresentati.

IX. Scalable oversight, debate e amplificazione iterativa

Il quinto limite strutturale del RLHF — la scalabilità del feedback umano con la capacità del modello — non è solo un problema pratico di costo del labeling: è un problema epistemico fondamentale. Quando un sistema di AI supera la competenza umana su un task specifico, i valutatori umani non sono in grado di distinguere le risposte corrette da quelle plausibili ma errate. Il problema della scalable oversight è quindi:

Silvano Salvador

come possiamo ottenere feedback allineato da valutatori umani su task che eccedono la loro competenza diretta?

Christiano, Leike, Brown e colleghi (2018) hanno proposto Iterated Amplification come risposta sistematica. L'idea fondamentale è decomporre i task complessi in sottotask che rimangono valutabili dagli umani, e usare un AI assistant per aiutare il valutatore umano a gestire questa decomposizione. Il ciclo di amplificazione iterativa è il seguente: al tempo t, il valutatore umano H^t con assistenza di un AI di livello t-1 produce segnali di training per un AI di livello t. Man mano che t aumenta, la complessità dei task gestibili cresce, ma la valutazione di ogni sottotask rimane alla portata di H con assistenza. La condizione di sicurezza dell'amplificazione iterativa è che l'AI al livello t non faccia nulla che H non avrebbe approvato se avesse avuto tempo illimitato e tutta l'assistenza necessaria.

Debate (Irving, Christiano e Amodei, 2018) adotta un approccio orthogonal: invece di amplificare il valutatore, si usa la struttura di un gioco a due giocatori per rendere più facile per il valutatore identificare le risposte errate. Due agenti AI argomentano posizioni opposte su una questione; il valutatore umano giudica quale argomento è più convincente. La tesi centrale del paper è che, anche se verificare una risposta è difficile, è spesso più facile per il valutatore identificare quando un argomento è fallace che costruire l'argomento corretto in autonomia. La struttura di gioco produce un equilibrio in cui l'agente veritiero — quello che argomenta per la risposta corretta — ha un vantaggio strutturale: ogni menzogna è vulnerabile alla confutazione da parte dell'avversario.

La formalizzazione in termini di teoria dei giochi è la seguente. Il debate è un gioco a somma zero tra agente A_1 (che argomenta la risposta r_1) e

Silvano Salvador

agente A_2 (che argomenta r_2), con payoff determinato dal giudizio del valutatore umano H. L'ipotesi fondamentale è che H sia un giudice polinomialmente limitato che può verificare prove di lunghezza polinomiale ma non generare dimostrazioni in autonomia. Sotto questa ipotesi, Irving e colleghi dimostrano che il gioco del debate ha un equilibrio in cui l'agente con la risposta corretta vince: se r_1 è vera e A_2 mente, A_1 può sempre presentare una confutazione che H può verificare in tempo polinomiale. La connessione con la teoria della complessità computazionale è esplicita: il risultato assume che il problema di verifica sia in NP (le prove si verificano in tempo polinomiale) anche se il problema di soluzione è intrattabile.

Reciprocal Reward Modeling (Leike et al., 2022) estende questi framework al caso in cui anche la valutazione debba essere appresa: si addestrano reward model multipli che si reciprocamente valutano, producendo un sistema di feedback distribuito che è meno vulnerabile al single point of failure del reward model singolo del RLHF standard. Process Reward Models (Lightman et al., 2023), introdotti da OpenAI per la matematica competitiva, spostano il feedback dalla risposta finale ai passi intermedi del ragionamento: invece di valutare se la soluzione è corretta, si valuta se ogni passo è logicamente valido. Questo rende la supervisione più scalabile perché verificare un singolo passo è computazionalmente più facile che verificare l'intera catena.

Il tema unificante di questi approcci — amplificazione, debate, process reward models — è che la supervisione scalabile non può essere solo una questione di raccogliere più feedback: deve modificare strutturalmente la relazione epistemica tra il valutatore umano e il sistema di AI, rendendo le risposte corrette più facili da verificare senza richiedere che il valutatore le generi. Questo è il correlato, nel dominio dell'allineamento, del principio che la compressione come intelligenza ha identificato come

Silvano Salvador

centrale nel Saggio II: la struttura del task di verifica deve essere sfruttata per ridurre la complessità del segnale di supervisione.

X. La correggibilità come proprietà strutturale: il framework di Russell

Stuart Russell, nel suo influente libro Human Compatible (2019) e nella serie di lavori tecnici che lo accompagnano, ha argomentato che il problema fondamentale dell'allineamento non è la progettazione di funzioni di ricompensa corrette ma la struttura epistemica dell'agente rispetto ai propri obiettivi. Un agente che massimizza una funzione di ricompensa fissa con certezza si comporta necessariamente in modo non correggibile — come mostrato dall'analisi della convergenza strumentale — mentre un agente che mantiene incertezza sulla propria funzione di ricompensa ha incentivi naturali a rimanere correggibile.

Il framework del Cooperative Inverse Reinforcement Learning (CIRL) formalizza questa intuizione. Si modella l'incertezza dell'agente sulla funzione di ricompensa come una distribuzione a priori P(R) sullo spazio delle possibili funzioni di ricompensa, aggiornabile bayesianamente sulla base del comportamento osservato dell'utente. L'agente massimizza l'utilità attesa rispetto a questa distribuzione:

$\pi^{*}_{CIRL}(s) = \text{argmax}_a \, \mathbb{E}_{R \sim P(R \mid \text{storia})} [\mathbb{E}_{\pi} [\sum_{t' \geq t} \gamma^{t'-t} R(s_{t'}, a_{t'}) \mid s_t = s, a_t = a]]$

dove P(R | storia) è la distribuzione aggiornata sulla funzione di ricompensa dopo aver osservato la storia delle preferenze dell'utente. L'aggiornamento bayesiano di P(R) in risposta alle osservazioni dell'utente è formalmente analogo al problema dell'IRL: ogni azione

Silvano Salvador

osservata dell'utente fornisce informazione su quale funzione di ricompensa sta massimizzando.

Un agente con questa struttura ha naturalmente un incentivo a raccogliere informazioni sul valore delle proprie azioni — chiedendo preferenze, osservando reazioni, permettendo correzioni — perché ogni nuova informazione riduce l'incertezza su R e aumenta la qualità dell'ottimizzazione a lungo termine. La correggibilità emerge come conseguenza dell'incertezza sugli obiettivi, non come proprietà imposta esternamente: un agente CIRL accetta le correzioni perché le tratta come informazione preziosissima sulla funzione di ricompensa vera.

La tensione tra correggibilità e ottimalità non si dissolve completamente, però. Un agente CIRL aggiorna P(R) bayesianamente ogni volta che osserva preferenze umane — ma se la distribuzione P(R) converge verso una funzione di ricompensa specifica con alta confidenza, la correggibilità si riduce: l'agente con alta confidenza nei propri obiettivi ha meno incentivo a sollecitare correzioni e più incentivo a resistere a modifiche che ridurrebbero la sua performance rispetto all'obiettivo attualmente creduto. Questo crea una dinamica preoccupante: man mano che un sistema di AI interagisce con l'utente e raccoglie più feedback, la sua stima degli obiettivi umani diventa più precisa, ma la sua disposizione alla correggibilità tende a ridursi — proprio quando la capacità di causare danni è al massimo.

XI. La verifica formale e i suoi limiti computazionali

La verifica formale dell'allineamento affronta un ostacolo computazionale che vale la pena rendere preciso. Un sistema di AI è sicuro rispetto a una proprietà Φ se per ogni input x nel dominio di

Silvano Salvador

deployment, l'output f(x) soddisfa $\Phi(f(x))$. La verifica di questa affermazione per una rete neurale profonda con n parametri e input continuo è un problema che appartiene alla classe PSPACE o superiore per proprietà generali. I verificatori neurali come Reluplex (Katz et al., 2017) e CROWN (Zhang et al., 2018) riducono il problema alla soddisfacibilità di sistemi lineari attraverso rilassamenti convessi delle non-linearità ReLU, producendo certificati di robustezza locali: garanzie che per ogni input x nella palla $l\infty$ di raggio ε attorno a un dato punto x_0, l'output non cambia label.

Questi certificati di robustezza locale non sono sufficienti per le proprietà di sicurezza globali rilevanti per l'allineamento — non fare mai X, garantire sempre Y — perché lo spazio degli input di un sistema di AI generale è non compatto e la proprietà di sicurezza deve valere uniformemente. La ricerca sul runtime monitoring affronta questo limite in modo pragmatico: invece di verificare le proprietà a priori, si monitora il comportamento in deployment e si interviene quando una proprietà sta per essere violata. Un monitor formale è un sistema M(x, f(x)) che emette un segnale di allarme prima che f(x) produca output pericolosi, permettendo all'operatore di intervenire o al sistema di modificare la propria azione. La limitazione è che il monitor deve essere più semplice e verificabile del sistema che monitora, il che crea un regresso: chi monitora il monitor?

La risposta più robusta al regresso del monitor è la separazione architetturale dei livelli di sicurezza. In un sistema a due moduli, un modulo principale M_1 (potente, non verificato) genera le azioni, mentre un modulo di sicurezza M_2 (più semplice, certificabile) filtra le azioni prima dell'esecuzione. La garanzia di sicurezza si sposta da M_1 a M_2: se M_2 è formalmente corretto rispetto alla proprietà di sicurezza, il sistema complessivo soddisfa quella proprietà indipendentemente dal

Silvano Salvador

comportamento di M_1. Questa architettura è implementata in alcune applicazioni safety-critical — il controllo di volo automatico con envelope protection, per esempio — ma la sua applicazione ai sistemi di AI generali richiede la capacità di specificare formalmente le proprietà di sicurezza rilevanti, il che ripropone il problema dell'allineamento a un livello di astrazione superiore.

Un approccio radicalmente diverso all'allineamento — concettualmente più vicino alla tradizione della verifica formale del software che al reinforcement learning — consiste nel tentativo di dimostrare formalmente che un sistema di AI soddisfa certe proprietà di sicurezza su tutti i possibili input. Se si potesse dimostrare con certezza formale che un sistema non produce mai output in una certa classe di risposte dannose, il problema dell'allineamento sarebbe almeno parzialmente risolto per quella classe.

La verifica formale di reti neurali è un campo attivo di ricerca sin dal lavoro fondazionale di Katz, Barrett, Dill, Julian e Kochenderfer (2017) con il framework Reluplex. Per reti neurali con attivazioni ReLU, il problema di verificare se l'output cade in una regione specificata per tutti gli input in un insieme dato è formulabile come un problema di programmazione lineare mista-intera (Mixed Integer Linear Programming, MILP): ogni neurone ReLU introduce una variabile binaria che indica se il neurone è attivo o inattivo, e la verifica richiede di esplorare tutte le possibili combinazioni di attivazione.

La complessità computazionale di questo problema è NP-completa nella dimensione della rete: il numero di stati ReLU da esplorare cresce esponenzialmente con il numero di neuroni. Per le reti usate nei sistemi di AI moderni — con miliardi di parametri e strati profondi — la verifica formale esatta è computazionalmente intrattabile. I metodi approssimati

Silvano Salvador

— come l'abstract interpretation applicata ai domain astratti degli intervalli (interval arithmetic), dei poliedri o dei zonotopi — rilassano il problema a uno computazionalmente trattabile, ma al costo di introdurre sovra-approssimazioni che possono portare a falsi negativi: certificano come non verificabili reti che sarebbero effettivamente sicure.

La limitazione più profonda della verifica formale per reti neurali ha radici nel primo saggio del pentattico, nella connessione con i risultati di indecidibilità di Turing. Per reti neurali ricorrenti (RNN) e per reti feedforward sufficientemente profonde con pesi sufficientemente generali, il problema di verificare se la rete produca mai un certo output su qualche input è formalmente equivalente al problema della fermata di macchine di Turing — e quindi indecidibile. Qualsiasi procedura di verifica automatica è pertanto necessariamente incompleta: o non termina su tutti gli input, o produce falsi positivi (certifica come sicuri sistemi che non lo sono), o produce falsi negativi. La terza impossibilità del pentattico risuona: la finitezza e la formalizzazione non possono catturare completamente la struttura infinita e aperta dei valori umani, e i limiti della verifica formale ne sono una manifestazione tecnica precisa.

IX-bis. Il reinforcement learning vincolato: sicurezza e MDP con costi

Il framework standard degli MDP massimizza un ritorno cumulativo scalare senza alcun vincolo esplicito sulla traiettoria. Per sistemi di AI che agiscono nel mondo reale, questa assenza di vincoli è una fonte diretta di pericolo: un agente ottimizzante che massimizza la ricompensa senza restrizioni può violare i vincoli fisici o di sicurezza che qualsiasi operatore razionale considererebbe non negoziabili. Il Constrained

Silvano Salvador

Markov Decision Process (CMDP) formalizza questo requisito aggiungendo al MDP standard un insieme di funzioni di costo $C_i : S \times A \to R$ e vincoli sulla loro aspettativa cumulativa scontata:

$$\max_\pi E_\pi[\sum_t \gamma^t R(s_t, a_t)] \quad \text{soggetto a: } E_\pi[\sum_t \gamma^t C_i(s_t, a_t)] \leq d_i \quad \text{per ogni } i = 1, ..., m$$

La struttura duale del CMDP ha una soluzione formale attraverso la teoria di Lagrange. Il lagrangiano del problema è:

$$L(\pi, \lambda) = E_\pi[\sum_t \gamma^t R(s_t, a_t)] - \sum_i \lambda_i (E_\pi[\sum_t \gamma^t C_i(s_t, a_t)] - d_i)$$

Il teorema di dualità forte garantisce che, sotto condizioni di regolarità (Slater constraint qualification), il punto di sella del lagrangiano coincide con la soluzione del problema primale vincolato. L'algoritmo Constrained Policy Optimization (CPO, Achiam et al., 2017) implementa questa struttura con aggiornamenti di policy gradient su un trust region che garantisce simultaneamente il miglioramento della ricompensa e il rispetto approssimativo dei vincoli di costo. La garanzia teorica di CPO è che, se la politica iniziale soddisfa i vincoli, ogni iterazione dell'algoritmo produce una politica che li soddisfa a meno di un errore controllabile dalla dimensione del trust region.

Proximal Policy Optimization Lagrangiana (PPO-Lagrangian, Ray et al., 2019) offre un'alternativa più semplice da implementare: si aggiornano simultaneamente i parametri della politica verso la ricompensa e i moltiplicatori di Lagrange λ_i verso i costi, con aggiornamenti opposti (gradient ascent su λ quando il vincolo è violato, gradient descent sulla politica). La convergenza è garantita verso un punto di sella sotto convessità del problema duale. In pratica, PPO-Lagrangiano produce comportamenti più sicuri della baseline non vincolata con degradazione moderata della performance.

Silvano Salvador

Per il problema dell'allineamento, i CMDP forniscono un inquadramento formale per la nozione intuitiva di "comportamento sicuro". Le funzioni di costo C_i codificano i vincoli di sicurezza — non fare X, non eccedere soglia Y, mantenere grandezza Z entro intervallo — separandoli dalla ricompensa principale. Questa separazione ha tre vantaggi strutturali rispetto all'approccio di penalizzazione scalare. Primo: i vincoli sono specificati in termini assolutamente, non come compromesso bilanciabile — la soglia d_i è un vincolo duro, non un peso nella funzione obiettivo. Secondo: i vincoli sono interpretabili e ispezionabili indipendentemente dalla ricompensa, facilitando l'auditing del comportamento. Terzo: la struttura lagrangiana produce un criterio di sicurezza che è robusto alla scala del problema: anche un agente molto capace è costretto a rispettare i vincoli se i moltiplicatori di Lagrange sono aggiornati appropriatamente.

Il limite principale dei CMDP è che i vincoli di costo devono essere specificati in anticipo come funzioni della coppia (stato, azione), il che ripropone il problema dell'allineamento a un livello diverso: come si specificano i vincoli in modo che catturino completamente ciò che si intende proibire? La risposta è che i CMDP non risolvono il problema dell'allineamento ma lo strutturano in modo più gestibile: separano la questione di "cosa si vuole ottenere" (ricompensa) da quella di "cosa non si vuole fare" (vincoli), e forniscono garanzie computazionali per il rispetto dei vincoli una volta specificati. Questa separazione corrisponde alla distinzione filosofica tra valori positivi (ciò che si persegue) e vincoli negativi (ciò che si evita) che le teorie etiche deontologiche hanno sempre considerato fondamentale.

Silvano Salvador

Dentro la scatola nera: interpretabilità meccanicistica e il programma aperto

I. Il problema dell'opacità come questione epistemica

Le parti precedenti di questo saggio hanno documentato le difficoltà strutturali dell'allineamento dal lato della specifica degli obiettivi: la legge di Goodhart, la convergenza strumentale, il teorema di Arrow. Esiste però una seconda dimensione del problema, distinta dalla prima ma intrecciata con essa in modo difficile da districare: l'opacità delle rappresentazioni interne dei sistemi neurali profondi. Anche se si disponesse di una funzione di ricompensa perfettamente allineata — circostanza che il teorema di Arrow rende concettualmente irraggiungibile nella sua forma più generale — resterebbe aperta la domanda su cosa fa concretamente il modello quando produce un output: quali strutture computazionali implementa, quali concetti rappresenta, quali meccanismi di ragionamento attiva. Senza risposta a questa domanda, la fiducia in un sistema di AI non può essere fondata sulla comprensione dei suoi meccanismi, ma soltanto sulla performance osservata in contesti noti — e la performance osservata può divergere dal comportamento in distribuzioni non viste in modi imprevedibili.

Il paradigma simbolico aveva, su questo punto, un vantaggio strutturale non trascurabile: un sistema esperto basato su regole esplicite è, per definizione, localmente interpretabile. Ogni inferenza può essere ricondotta alla catena di regole che l'ha prodotta; ogni conclusione può essere giustificata in termini di premesse verificabili; ogni comportamento inatteso può essere tracciato alla regola responsabile. La

Silvano Salvador

trasparenza è una proprietà architettonica del sistema, non un risultato da conquistare a posteriori. Il prezzo — l'incapacità di operare in domini ad alta complessità statistica, documentata nel primo saggio — è enorme, ma la trasparenza è reale.

I sistemi neurali profondi invertono questo compromesso: sono capaci nei domini complessi ma opachi nei propri meccanismi interni. La ricerca sull'interpretabilità meccanicistica — il programma che studia la struttura computazionale interna delle reti neurali — è il tentativo di recuperare comprensione senza sacrificare capacità. Non si tratta di ritornare al simbolismo: le capacità dei modelli neurali dipendono dalla struttura distribuita e continua delle loro rappresentazioni, che non può essere ridotta a manipolazione simbolica senza perdere le proprietà essenziali che la rendono potente. Si tratta invece di analizzare retrospettivamente i meccanismi che si sono sviluppati durante l'addestramento, cercando di tradurli in strutture comprensibili — un'operazione che combina algebra lineare, geometria differenziale, teoria dell'informazione e analisi causale in modi che il campo sta ancora esplorando.

II. La struttura delle rappresentazioni neurali: dalla tabella alla geometria

La prima questione dell'interpretabilità meccanicistica è la struttura delle rappresentazioni intermedie. In una rete neurale profonda con L strati, lo strato l mappa l'input attraverso una sequenza di trasformazioni non lineari, producendo vettori di attivazione h^l(x) $\in \mathbb{R}^{d_l}$ per ogni input x. Questi vettori vivono in uno spazio ad alta dimensionalità con geometria non banale, e la domanda è quale struttura abbiano — se le singole coordinate corrispondano a concetti identificabili, se esistano

Silvano Salvador

direzioni interpretabili nello spazio di attivazione, se la geometria dello spazio rispecchi la struttura semantica del dominio.

La risposta più semplice — che le coordinate individuali corrispondano a concetti identificabili, con il neurone i attivo per il concetto X e il neurone j per il concetto Y — è smentita dall'evidenza empirica sistematica. I vettori di attivazione non hanno tipicamente struttura assiale: non esiste un allineamento tra l'asse coordinato i e un concetto semantico identificabile. La rappresentazione è distribuita: ogni concetto è codificato da un pattern di attivazione su molti neuroni, e ogni neurone contribuisce a molti concetti diversi. Questa struttura distribuita è la manifestazione computazionale di ciò che il secondo saggio ha descritto come rappresentazione tramite distribuzione latente — la conoscenza del modello è distribuita nei pesi, non localizzata in componenti identificabili.

Ma la struttura distribuita non implica assenza di struttura. I vettori di attivazione di un modello addestrato su un corpus linguistico mostrano regolarità geometriche notevoli. La proprietà di linearità delle relazioni semantiche, scoperta da Mikolov e colleghi nel 2013 nei modelli word2vec, è l'esempio più famoso: il vettore della differenza v("re") − v("uomo") + v("donna") è vicino a v("regina") nello spazio delle rappresentazioni. Questa regolarità suggerisce che le relazioni semantiche siano codificate come vettori di traslazione nello spazio di attivazione — una struttura algebrica lineare che permette di ragionare sulle relazioni semantiche attraverso operazioni vettoriali.

III. La superposition hypothesis e la geometria delle feature

Silvano Salvador

Un risultato fondamentale dell'interpretabilità meccanicistica, derivato sistematicamente da Elhage, Nanda, Olsson e colleghi di Anthropic nel 2022, è la superposition hypothesis: le reti neurali rappresentano più feature di quante abbiano neuroni, codificando feature diverse in direzioni quasi-ortogonali dello spazio di attivazione. Se il modello ha n neuroni in un dato strato ma deve rappresentare m > n concetti distinti, esso può codificare i concetti come vettori unitari in $\mathbb{R}^n$ che, pur non essendo esattamente ortogonali, sono sufficientemente distanti angolarmente da essere distinguibili con alta probabilità per qualsiasi coppia di feature.

La struttura geometrica ottimale di questa rappresentazione è quella di un frame equangolare stretto (tight equiangular frame): una famiglia di m vettori unitari $v_1, ..., v_m \in \mathbb{R}^n$ con m > n che minimizzano la massima interferenza mutua, cioè minimizzano $\max_\{i\neq j\} |\langle v_i, v \rangle|$. Per $m \leq n^2$, esiste sempre un frame equiangolare stretto con $|\langle v_i, v \rangle| = \sqrt{((m-n)/(n(m-1)))}$, e queste strutture sono realizzate da famiglie di vettori come i simplessi regolari, i frame di Gabor e le configurazioni di linee equiangolari in geometria combinatoria.

La trade-off fondamentale della superposition è tra la capacità (quante feature si possono rappresentare) e la precisione (quanto accuratamente si può distinguere tra feature diverse). Con n neuroni, si possono rappresentare esattamente n feature in modo perfettamente ortogonale, oppure $O(n^2)$ feature con interferenza proporzionale a $1/\sqrt{n}$, oppure ancora più feature accettando interferenze maggiori. La rete sceglie implicitamente questo punto nel trade-off durante l'addestramento, e la struttura ottimale dipende dalla frequenza relativa delle feature nel corpus di addestramento — feature rare richiedono meno precisione e possono essere compresse più aggressivamente.

Silvano Salvador

La superposition ha una conseguenza importante per l'interpretabilità: le feature del modello non sono le coordinate dello spazio di attivazione ma direzioni arbitrarie in tale spazio. Identificare queste direzioni richiede tecniche di decomposizione sparse, tipicamente formulate come il seguente problema di ottimizzazione:

$$\min_{D \in \mathbb{R}^{d\times k}, C = \{c_i\}} \sum_i \|h_i - D\, c_i\|^2 + \lambda \|c_i\|_1 \text{ s.t. } \|D_j\| = 1 \; \forall j$$

dove $h_i \in \mathbb{R}^d$ è il vettore di attivazione dell'i-esimo input, $D \in \mathbb{R}^{d\times k}$ è il dizionario con $k > d$ atomi normalizzati, $c_i \in \mathbb{R}^k$ è il codice sparse dell'i-esimo input, e il termine di penalità L_1 $\|c_i\|_1$ promuove la sparsità — pochi atomi del dizionario attivi per ogni input. Le colonne di D così apprese — le sparse features o dictionary atoms — sono le candidate rappresentazioni interpretabili: direzioni nello spazio di attivazione che corrispondono a concetti identificabili nel dominio.

Il metodo delle Sparse Autoencoders (SAE), sviluppato sistematicamente da Cunningham, Ewart, Riggs, Huben e Sharkey (2023) e poi esteso da Templeton e colleghi di Anthropic nel 2024, applica questa decomposizione a scala su grandi modelli linguistici come Claude. Una SAE è un autoencoder con un collo di bottiglia espansivo: l'encoder proietta le attivazioni h da $\mathbb{R}^d$ a $\mathbb{R}^k$ con $k \gg d$, il bottleneck applica una non-linearità ReLU per forzare sparsità, e il decoder ricostruisce le attivazioni originali. L'addestramento minimizza la perdita di ricostruzione con penalità L_1 sulla sparsità:

$$L(W_{enc}, W_{dec}) = \|h - W_{dec} \cdot \mathrm{ReLU}(W_{enc}\, h + b_{enc})\|^2 + \lambda \|\mathrm{ReLU}(W_{enc}\, h + b_{enc})\|_1$$

Le colonne di W_dec corrispondono alle direzioni delle sparse features nello spazio di attivazione originale. In applicazioni recenti a modelli come GPT-4 e Claude, questo metodo ha identificato milioni di feature

Silvano Salvador

interpretabili che corrispondono a concetti come "codice Python", "nomi propri femminili", "espressioni di incertezza", "testi legali", con una granularità e una chiarezza semantica sorprendenti.

IV. I circuiti come unità computazionali identificabili

Oltre alle rappresentazioni, l'interpretabilità meccanicistica studia i circuiti: sottografi del grafo computazionale della rete che implementano funzioni identificabili. L'idea che le reti neurali implementino algoritmi specifici attraverso la composizione di operazioni elementari localizzate in sottoinsiemi identificabili di pesi è emersa dal lavoro pioneristico di Olah, Satoshi Cammarata e colleghi su reti convoluzionali per il riconoscimento di immagini (2020). In questo lavoro sono stati identificati circuiti che implementano il rilevamento di curve attraverso la composizione di rilevatori di bordo orientati; circuiti per il rilevamento di frequenze spaziali; circuiti per la rappresentazione di texture specifiche. Queste strutture mostrano una corrispondenza notevole con le risposte dei neuroni nelle aree V1 e V2 della corteccia visiva primaria dei primati — una convergenza evolutiva e computazionale tra sistemi biologici e artificiali ottimizzati sullo stesso tipo di dati.

Per i modelli linguistici, Wang, Variengien, Conmy, Shlegeris e Steinhardt (2022) hanno identificato nel GPT-2 un circuito specifico per il completamento di sequenze indirette di oggetti — compiti del tipo "Maria ha dato il libro a Giovanni. Giovanni ha detto ad Alice che ..." (risposta attesa: "il libro era di Maria") — che implementa una forma di ragionamento transitivo attraverso la composizione di teste di attenzione con funzioni di copia, recupero e mediazione. Il circuito identificato coinvolge circa trenta teste di attenzione su ventiquattro strati, e può

Silvano Salvador

essere analizzato componente per componente per comprendere il contributo di ciascuna testa.

La metodologia per l'identificazione di circuiti si articola in tre fasi. La prima è l'activation patching o causal tracing (Meng, Geva, Belinkov, Berant, Bau, 2022): si interviene direttamente sulle attivazioni sostituendo il vettore di attivazione di un componente (un neurone, una testa di attenzione, uno strato) con quello calcolato su un input di riferimento diverso, e si misura l'effetto sull'output. Se la sostituzione cambia l'output del modello in modo coerente con l'input di riferimento, quel componente è causalmente rilevante per il comportamento in esame. Formalmente, si definisce l'effetto causale indiretto del componente k sull'output come:

$$CIE(k) = \mathbb{E}_{x, x_{ref}} [f(x, do(h^k \leftarrow h^k(x_{ref}))) - f(x)]$$

dove $do(h^k \leftarrow v)$ indica l'intervento che sostituisce le attivazioni del componente k con il valore v, mantenendo fissi tutti gli altri input. La notazione $do(\cdot)$ è quella della causalità di Pearl, che permette di distinguere interventi (modifiche dirette) da osservazioni (condizionamento su variabili osservate), evitando confusione tra correlazione e causalità.

La seconda fase è la component attribution: dati i componenti causalmente rilevanti identificati dall'activation patching, si analizza il loro contributo funzionale attraverso l'analisi dello spazio delle query-key nei transformer (quali posizioni vengono messe in relazione dall'attenzione), dello spazio dei valori (quali informazioni vengono copiate o trasformate) e della struttura delle connessioni tra strati. La terza fase è la verifica per ablazione: si disabilita sistematicamente ciascun componente del circuito candidato — sostituendone le attivazioni con zero o con la media sulle attivazioni del corpus — e si verifica che il

Silvano Salvador

comportamento del modello degradi in modo proporzionale all'importanza attribuita al componente.

V. Probing, causalità e dormant coding

Un approccio complementare all'analisi dei circuiti è il probing: addestrare classificatori lineari leggeri sulle rappresentazioni intermedie della rete per verificare se certe proprietà del dominio siano linearmente decodificabili. Se un probing lineare — un classificatore della forma f(h) = w^T h + b — classifica accuratamente la funzione sintattica delle parole a partire dalle attivazioni di uno strato, si conclude che quello strato codifica linearmente l'informazione sintattica.

Il probing è metodologicamente semplice ma epistemicamente limitato: esso fornisce evidenze correlazionali — dimostra che l'informazione sulla proprietà è presente nelle attivazioni, geometricamente in forma lineare — ma non fornisce evidenze causali che questa informazione sia effettivamente usata dal modello per calcolare l'output. Il fenomeno del dormant coding illustra questa distinzione: un modello potrebbe codificare linearmente la funzione sintattica nelle attivazioni dello strato l senza mai utilizzare questa rappresentazione negli strati successivi, per esempio perché l'informazione è stata codificata da meccanismi ausiliari durante l'addestramento ma non contribuisce al calcolo del percorso principale.

Per stabilire la causalità — verificare che la rappresentazione di una proprietà nello strato l influenzi causalmente l'output del modello — sono necessari metodi interventistici. L'activation patching già descritto è uno di questi; la Linear Representation Probing causale (Geiger, Lu, Icard, Potts, 2021) è un altro: si identifica una direzione lineare nel spazio di attivazione che codifica una proprietà di interesse, poi si interviene

Silvano Salvador

ruotando i vettori di attivazione lungo questa direzione mentre si mantengono ortogonali gli altri componenti, e si verifica se l'output cambia in modo coerente con la perturbazione introdotta.

La struttura formale dell'analisi causale applicata alle reti neurali si fonda sul formalismo degli interventi di Pearl (2000). Un circuito computazionale in una rete neurale può essere modellato come un DAG causale $G = (V, E)$, dove V è l'insieme dei nodi (input, attivazioni intermedie, output) e E è l'insieme degli archi (connessioni tra nodi). La distribuzione di probabilità $P(V)$ fattorizza secondo il DAG, e gli interventi $do(X_i = x_i)$ — sostituire il valore del nodo X_i con x_i indipendentemente dai suoi genitori — producono una distribuzione intervenuta $P^{do(X_i = x_i)}(V)$ che si può calcolare tramite il calcolo do, eliminando i termini $P(X_i \mid pa(X_i))$ dalla fattorizzazione e sostituendo con $\delta(X_i - x_i)$. L'effetto causale del componente X_i sull'output Y è:

$$ACE(X_i \to Y) = \sum_y y \, [P^{do(X_i = 1)}(Y = y) - P^{do(X_i = 0)}(Y = y)]$$

e la sua identificabilità dal DAG causale segue dai criteri di backdoor e frontdoor di Pearl. La trasposizione di questo formalismo alle reti neurali richiede di modellare le teste di attenzione e i layer FFN come nodi del DAG, il che è computazionalmente costoso per modelli di dimensioni frontier ma praticamente realizzabile per modelli di dimensioni moderate.

VI. L'interpretabilità meccanicistica come programma di sicurezza

L'interpretabilità meccanicistica non è soltanto una disciplina con interesse scientifico intrinseco: è, nelle intenzioni di una parte

Silvano Salvador

significativa dei suoi praticanti, un programma di sicurezza con motivazioni specifiche legate al problema dell'allineamento. L'argomentazione fondamentale è che un sistema di AI di cui si comprende il meccanismo computazionale interno è più sicuro di uno di cui si comprende soltanto il comportamento osservato, per almeno tre ragioni distinte.

La prima ragione è l'anticipazione dei comportamenti fuori distribuzione. Un modello il cui meccanismo è compreso a livello di circuiti permette di ragionare su come si comporterà in situazioni non viste nell'addestramento, perché si può analizzare se i circuiti identificati generalizzano in modo coerente o se presentano singolarità che potrebbero manifestarsi in distribuzioni diverse. Un modello di cui si conosce soltanto la performance su benchmark standard non fornisce garanzie analoghe.

La seconda ragione è l'identificazione di meccanismi di fallimento prima della manifestazione. L'analisi delle rappresentazioni interne può rivelare la presenza di concetti o associazioni potenzialmente problematiche che non emergono nel comportamento osservato in condizioni normali ma potrebbero manifestarsi in condizioni particolari. Per esempio, l'identificazione di feature che codificano associazioni stereotipate su gruppi sociali può anticipare comportamenti discriminatori anche se il modello non li manifesta nei benchmark standard, perché le feature esistono nelle rappresentazioni indipendentemente da se vengono attivate nei benchmark.

La terza ragione è la progettazione di interventi mirati. La comprensione meccanicistica permette di identificare con precisione quali componenti della rete sono responsabili di comportamenti indesiderati, e di intervenire chirurgicamente su quei componenti senza degradare le

Silvano Salvador

capacità non correlate. Il paradigma alternativo — il fine-tuning generico su dataset di allineamento — può migliorare il comportamento osservato nei benchmark ma può anche modificare in modo imprevedibile le capacità del modello in domini non monitorati.

VII. AIXI, la convergenza strumentale e il limite teorico dell'AGI

Il problema dell'allineamento assume la sua forma più acuta — e più speculativa — quando si considera il concetto di intelligenza artificiale generale (AGI): un sistema capace di apprendere e ottimizzare in qualsiasi dominio con competenza paragonabile o superiore a quella umana. Il limite teorico dell'AGI è fornito da AIXI, la costruzione di Marcus Hutter (2000, 2005) come formalizzazione dell'agente universalmente ottimale.

AIXI è un agente che, a ogni passo temporale t, seleziona l'azione che massimizza il ritorno atteso secondo il prior di Solomonoff sulla distribuzione dell'ambiente — la distribuzione di probabilità che assegna a ogni possibile modello computazionale dell'ambiente una probabilità proporzionale a 2^{−K(μ)}, dove K(μ) è la complessità di Kolmogorov del modello μ (la lunghezza del programma più corto che lo implementa su una macchina universale di Turing prefissa U). L'azione ottimale di AIXI è:

a*$_t$ = *argmax*{a_t} ∑{*o_t r_t*} · *max*{a_{t+1}} ∑{*o*{t+1} r_{t+1}} · ... · max_{a_m} ∑{*o_m r_m*} *(r_t* + *r*{t+1} + ... + r_m) · ∑{*q : U(q, a*{1:t} o_{1:t-1}) = o_{1:t} r_{1:t}} 2^{-|q|}

dove la somma finale è sul peso di Solomonoff di tutti i programmi q consistenti con le osservazioni e le ricompense passate. AIXI agisce

Silvano Salvador

come se possedesse un prior bayesiano universale su tutti i possibili ambienti computabili, e lo aggiornasse bayesianamente sulla base delle osservazioni — massimizzando poi il ritorno atteso rispetto alla distribuzione a posteriori.

Il teorema fondamentale di AIXI afferma che la perdita cumulativa di AIXI rispetto a un agente che conosce il vero modello dell'ambiente μ è limitata superiormente da $K(\mu) / \log(1/\gamma)$, dove $K(\mu)$ è la complessità di Kolmogorov del vero ambiente e γ è il fattore di sconto. Per ambienti con struttura regolare — piccola complessità di Kolmogorov — AIXI apprende rapidamente e il regret cumulativo è piccolo; per ambienti complessi il regret è maggiore ma rimane limitato.

AIXI è però computazionalmente inaccessibile nella sua forma esatta: il calcolo dell'azione ottimale richiede la somma su tutti i programmi di Turing di lunghezza arbitraria che sono consistenti con le osservazioni — un'operazione che, essendo correlata alla funzione di occupazione delle macchine di Turing, non è calcolabile. È un limite asintotico teorico, non un'architettura implementabile. La distanza tra i sistemi di AI pratici e il limite di AIXI è la distanza tra l'approssimabile e l'ottimale, una distanza che le leggi di scala mostrano decrescere regolarmente con le risorse computazionali senza mai annullarsi in un sistema finito.

Per il problema dell'allineamento, AIXI ha un'implicazione strutturale che riprende con forza i temi precedenti: un sistema che si avvicina all'ottimalità universale ha, per costruzione, tutte le proprietà di convergenza strumentale descritte nella parte precedente. Un AIXI-approssimato sufficientemente capace cercherà di preservare i propri obiettivi, acquisire risorse e resistere alla correzione — non perché sia "malvagio" nel senso morale, ma perché queste sono le strategie strumentalmente ottimali per qualsiasi funzione di ricompensa non

Silvano Salvador

banale. L'allineamento di un sistema vicino al limite di AIXI richiede che la funzione di ricompensa stessa sia allineata con i valori umani — e il teorema di Arrow afferma che questa allineamento non può essere raggiunto mediante una funzione scalare che soddisfi requisiti minimi di razionalità collettiva. Le tre impossibilità convergono nuovamente verso la stessa struttura.

Conclusione — Tre impossibilità, un principio

I. Il filo unificatore

Questi primi tre saggi hanno percorso un arco che copre tre secoli di matematica — dalla logica di Leibniz ai sistemi formali di Hilbert, dalla teoria dell'informazione di Shannon alla geometria riemanniana delle distribuzioni, dalla teoria della scelta sociale di Arrow alla teoria dei giochi di Nash — e tre decenni di storia dell'intelligenza artificiale, dalla crisi del paradigma simbolico al reinforcement learning profondo e al problema dell'allineamento. La struttura sottostante a questo percorso non è enciclopedica ma analitica: ogni saggio sviluppa una variazione dello stesso tema fondamentale, con crescente concretezza e urgenza pratica.

Il tema è quello dell'impossibilità: la dimostrazione, in tre domini distinti ma analogamente strutturati, che la formalizzazione finita non può catturare completamente una struttura che è costitutivamente infinita e aperta. Nel primo saggio, la struttura infinita era la matematica stessa — i teoremi di Gödel dimostrano che in ogni sistema formale

Silvano Salvador

sufficientemente espressivo esistono proposizioni vere ma indimostribili, e che il sistema non può dimostrare la propria coerenza dall'interno. Nel secondo saggio, la struttura infinita era la distribuzione generativa del mondo — nessun modello parametrico finito può catturare completamente la distribuzione che genera i dati reali, e la distanza dal prior universale di Solomonoff non si annulla mai. Nel terzo saggio, la struttura infinita è l'insieme dei valori umani — il teorema di Arrow afferma che non esiste funzione di benessere sociale scalare che li aggreghi coerentemente, e la legge di Goodhart afferma che qualsiasi proxy scalare usata come obiettivo viene sfruttata in modi che tradiscono l'intenzione originale.

Il principio generale che le unifica — e che questi tre saggi propongono come filo conduttore per comprendere i fondamenti dell'intelligenza artificiale — può essere enunciato così: ogni sistema formale finito, applicato a una struttura che lo trascende per apertura e complessità, produce risultati parzialmente corretti e vuoti irriducibili che non possono essere colmati dall'interno del sistema stesso. È un principio che attraversa la matematica (Gödel, Turing), la teoria dell'informazione (Shannon, Solomonoff), la teoria della scelta sociale (Arrow, Sen) e la teoria della decisione (Goodhart, Omohundro). Nelle tre variazioni dei rispettivi saggi, i sistemi formali finiti sono rispettivamente: i sistemi formali di Hilbert, i modelli parametrici finiti, le funzioni di ricompensa scalari.

II. Il ruolo epistemico della matematica

Attraverso i tre saggi è emerso un quadro coerente del ruolo della matematica nell'intelligenza artificiale che differisce sia dalla visione del paradigma simbolico sia da quella del pragmatismo ingenuo che a volte

Silvano Salvador

accompagna l'apprendimento automatico applicato. Il paradigma simbolico vedeva la logica formale come il fondamento dell'intelligenza: l'intelligenza è computazione logica, e la matematica ne fornisce la struttura completa. Il pragmatismo ingenuo vede la matematica come strumento: quello che conta è la performance empirica, e la matematica serve a descrivere e migliorare i sistemi.

La posizione che abbiamo sviluppato è diversa da entrambe. La matematica non è né il fondamento né soltanto lo strumento dell'intelligenza: è il mezzo attraverso cui le proprietà profonde dei sistemi cognitivi diventano visibili con precisione sufficiente da poter essere discusse con rigore e produttività. I teoremi di incompletezza di Gödel non fondano l'intelligenza ma ne rivelano i limiti strutturali; la teoria di Shannon non fonda la cognizione ma rivela che la cognizione è essenzialmente gestione dell'incertezza in un senso matematicamente preciso; la teoria dei giochi non fonda l'agire razionale ma rivela le strutture di incompatibilità tra razionalità individuale e collettiva. La matematica, in questo senso, è uno strumento di chiarificazione concettuale di straordinaria potenza — più potente, in molti contesti, del linguaggio naturale o della riflessione filosofica informale, perché costringe alla precisione delle premesse e rende impossibile nascondere le assunzioni cruciali.

Un terzo livello merita di essere reso esplicito, al di là del fondazionalismo e dello strumentalismo. La matematica ha svolto in questi saggi un ruolo — che si potrebbe chiamare diagnostico — che non è riducibile ai primi due. Le strutture matematiche hanno permesso di «vedere» i limiti dei sistemi prima che questi si manifestassero empiricamente, identificando dove e perché un approccio sarebbe fallito. Il teorema di Arrow non solo descrive perché l'aggregazione delle preferenze è difficile: diagnostica la struttura esatta della difficoltà e

Silvano Salvador

indica dove cercare soluzioni parziali. La non-calcolabilità del prior di Solomonoff non solo descrive un limite computazionale: indica che qualsiasi proxy calcolabile avrà un errore irriducibile e di quale tipo. Questa funzione diagnostica della matematica — produrre mappe dei limiti prima dell'esplorazione empirica — è forse il suo contributo più prezioso all'intelligenza artificiale come disciplina scientifica.

La disputa tra formalisti e intuizionisti nella fondazione della matematica del primo Novecento aveva già identificato questa dualità: la matematica formale produce risultati che non erano accessibili all'intuizione, ma l'intuizione seleziona quali formalismi sono matematicamente interessanti. Nell'intelligenza artificiale la dualità si ripresenta: i risultati empirici selezionano quali strutture matematiche sono rilevanti (le leggi di scala non erano attese dai teorici dell'informazione prima che emergessero dai dati), mentre la struttura matematica anticipa i fallimenti che l'empiria scoprirà più tardi (la legge di Goodhart era nota prima che il reward hacking nei sistemi moderni la confermasse). Il dialogo tra teoria e pratica nell'IA è quindi un caso particolare del dialogo tra deduzione e induzione che la filosofia della scienza ha studiato per secoli, con la specificità che la velocità del progresso empirico rende la funzione anticipatoria della matematica particolarmente preziosa.

Il fallimento del paradigma simbolico era, in parte, un fallimento matematico: le strutture logiche usate come fondamento erano inadeguate alla natura aperta dei problemi affrontati, e la matematica ha reso visibile questa inadeguatezza attraverso risultati come il problema del frame e la texture aperta dei concetti. Il successo del paradigma statistico è, in parte, un successo matematico: la teoria dell'informazione, la geometria riemanniana delle distribuzioni, la teoria dei punti fissi e la programmazione dinamica hanno fornito strumenti adeguati alla

Silvano Salvador

struttura dei problemi, e i risultati empirici rispecchiano questa adeguatezza strutturale.

III. Le tre impossibilità come programma positivo

Le tre impossibilità documentate finora nei nostri contributi possono sembrare, a una prima lettura, il fondamento di un pessimismo sistematico sull'intelligenza artificiale: i sistemi simbolici non possono catturare l'intelligenza completa, i sistemi statistici non possono approssimare perfettamente la distribuzione del mondo, i sistemi di allineamento non possono specificare completamente i valori umani. Questa lettura fraintende la funzione dei teoremi di impossibilità in matematica, e il trio di saggi ha sviluppato in modo ricorrente la lettura alternativa.

Il teorema di incompletezza di Gödel non ha paralizzato la matematica: ha liberato la ricerca fondazionale dalla ricerca di un fondamento che non può esistere, aprendo la strada a un pluralismo fondazionale fecondo — la reverse mathematics, la teoria dei grandi cardinali, la homotopy type theory — e ha mostrato che la ricchezza della matematica è inesauribile proprio perché nessun sistema formale la cattura nella sua totalità. Analogamente, l'impossibilità di catturare la distribuzione del mondo con un modello parametrico finito non ha paralizzato l'apprendimento automatico: ha mostrato che il programma positivo è l'avvicinamento asintotico, regolato dalle leggi di scala, al limite irraggiungibile del prior di Solomonoff. E l'impossibilità di Arrow non paralizza il problema dell'allineamento: mostra che il programma positivo deve abbandonare la ricerca di soluzioni perfette — funzioni di ricompensa scalari che catturino completamente i valori collettivi — e

Silvano Salvador

sviluppare invece approcci multi-livello che combinino feedback iterativo, incertezza esplicita sugli obiettivi, interpretabilità meccanicistica e strutture istituzionali per la governance distribuita.

I sistemi di intelligenza artificiale che oggi costituiscono lo stato dell'arte — GPT-4 (OpenAI, 2023), Claude 3 (Anthropic, 2024), Gemini Ultra (Google DeepMind, 2024) — sono implementazioni concrete delle strutture matematiche sviluppate in questo saggio. Sono transformer autoregressivi di grandissima scala addestrati con RLHF o con varianti come Constitutional AI e DPO; esibiscono le proprietà di emergenza descritte dalle leggi di scala; soffrono delle fragilità strutturali del RLHF identificate nel capitolo precedente; e sono soggetti alla legge di Goodhart in ogni deployment in cui l'obiettivo di addestramento diverge dall'obiettivo del deployment. L'analisi formale di questo saggio non descrive sistemi ipotetici futuri ma i sistemi che esistono oggi e che vengono usati da centinaia di milioni di persone — con le loro capacità straordinarie e con le loro impossibilità strutturali egualmente reali.

Il programma positivo nell'allineamento si articola in quattro direzioni complementari che la ricerca sta esplorando con intensità crescente. La prima è il miglioramento dei meccanismi di feedback: sviluppare metodi di elicitazione delle preferenze che catturino in modo più fedele le preferenze reali degli utenti diversi, riducendo i bias del campionamento demografico e la non-stazionarietà contestuale. La seconda è la rappresentazione esplicita dell'incertezza sugli obiettivi: costruire agenti che mantengano distribuzioni su possibili funzioni di ricompensa e agiscano in modo correggibile in risposta a questa incertezza, come nel framework CIRL. La terza è l'interpretabilità meccanicistica scalabile: sviluppare metodi automatici per identificare le rappresentazioni e i circuiti che implementano comportamenti potenzialmente problematici, permettendo interventi mirati prima che i comportamenti si manifestino.

Silvano Salvador

La quarta è la governance istituzionale: progettare strutture organizzative e regolamentari che distribuiscano il potere di influenzare il comportamento dei sistemi di AI tra gruppi diversificati di portatori di interessi, riducendo il rischio di dittatura preferenziale.

III-bis. Verso la formalizzazione dell'analogia: struttura comune e limiti della trasposizione

Le tre impossibilità che attraversano i nostri papers sono state presentate come variazioni di un medesimo tema. Vale però la pena chiarire con precisione cosa questo accostamento significhi formalmente, distinguendo ciò che è dimostrabile da ciò che è congettura filosofica produttiva ma non ancora teorema.

La struttura condivisa può essere articolata in termini di una tripla (F, Σ, C), dove F è un sistema formale finito, Σ è una struttura che F tenta di catturare o approssimare, e C è un criterio di adeguatezza. In ciascuna delle tre impossibilità, questa tripla assume una forma specifica. Nel caso di Gödel: F è un sistema assiomatico ricorsivamente enumerabile sufficientemente espressivo (estende l'aritmetica di Peano), Σ è l'insieme delle verità aritmetiche, C è la derivabilità $\vdash_F \sigma$. Il teorema di incompletezza afferma che $\exists\ \sigma \in \Sigma$ tale che $\neg\vdash_F \sigma \wedge \neg\vdash_F \neg\sigma$ (Gödel, 1931). Nel caso di Solomonoff: F è un modello parametrico M_θ con capacità finita, Σ è la distribuzione generativa vera del dominio approssimata dal prior universale P* di Solomonoff, C è la divergenza KL. Il teorema di dominanza universale afferma che $D_{KL}(P^* \| M_\theta) > 0$ per ogni M_θ di dimensione finita (poiché P* non è calcolabile e M_θ lo è). Nel caso di Arrow-Goodhart: F è una funzione scalare $R: S\times A \to \mathbb{R}$ come proxy degli obiettivi, Σ è l'insieme dei valori umani U, C è il misalignment gap $\Delta(\pi^*)$. Il teorema di Goodhart afferma che per

Silvano Salvador

ogni R correlata ma non identica a U, $\Delta(\pi^*) > 0$ sotto ottimizzazione sufficientemente capace.

La struttura condivisa è la seguente: in ciascun caso, l'adeguatezza di F rispetto a Σ sotto C è non raggiungibile dall'interno di F, nel senso che la struttura di Σ eccede ciò che F può rappresentare o computare nei propri termini. Per Gödel, la verità aritmetica non è coestensiva con la derivabilità in nessun sistema formale ricorsivo: c'è sempre un divario tra Th(ℕ) e {Th_F : F ricorsivo, consistente}. Per Solomonoff, la complessità di Kolmogorov non è calcolabile, e qualsiasi proxy calcolabile è dominata da P* con costante di approssimazione che non si annulla mai. Per Goodhart, la proiezione dei valori umani su uno scalare perde irriducibilmente le dimensioni di variazione contestuale e la non-stazionarietà che il teorema di Arrow formalizza.

Tre precisazioni sono però essenziali per non confondere l'analogia con un'unità formale. Prima: la natura del divario è diversa nelle tre impossibilità. Nel caso di Gödel il divario è logico — una proposizione vera indimostrabile — e si manifesta in modo costruttivo attraverso la formula che asserisce la propria indimostrabilità. Nel caso di Solomonoff il divario è computazionale — la non-calcolabilità del prior universale — e si manifesta asintoticamente nella curva delle leggi di scala. Nel caso di Arrow-Goodhart il divario è game-teorico — l'ottimizzazione sfrutta la non-identità tra proxy e obiettivo — e si manifesta empiricamente nello specification gaming. Questi sono tre tipi di impossibilità distinti, collegati da analogia strutturale ma non deducibili l'uno dall'altro.

Seconda precisazione: il teorema di Arrow, nella sua formulazione canonica, riguarda l'aggregazione di preferenze ordinali sotto condizioni di razionalità collettiva, non la rappresentazione di valori in funzioni scalari continue. Il collegamento con la legge di Goodhart richiede una

Silvano Salvador

traduzione: le "alternative" di Arrow diventano le "politiche" dell'agente; gli "individui" diventano i "contesti e valutatori"; la "funzione di benessere sociale" diventa il "reward model". Questa traduzione è concettualmente plausibile — e il quarto limite del RLHF discusso nel capitolo precedente la istanzia concretamente — ma non è un'applicazione letterale del teorema di Arrow all'ambito del reinforcement learning.

Terza precisazione: l'impossibilità di Gödel è assoluta nel senso logico — non dipende dalla capacità computazionale del sistema formale, ma dalla sua sola capacità di rappresentare l'aritmetica. Le impossibilità di Solomonoff e di Goodhart sono invece condizionali — dipendono rispettivamente dalla finitezza del modello e dalla capacità dell'agente ottimizzante. Questo suggerisce che la terza impossibilità è mitigabile in linea di principio attraverso modelli sufficientemente grandi (Solomonoff) o funzioni di ricompensa sufficientemente ricche e aggiornate (Goodhart), mentre l'incompletezza gödeliana non lo è. L'analogia è illuminante come mappa delle strutture, non come prova che il problema dell'allineamento sia irrisolvibile nel senso in cui lo è l'incompletezza aritmetica.

Detto questo, il principio che i nostri tre papers propongono come filo conduttore — ogni sistema formale finito applicato a una struttura che lo trascende per apertura e complessità produce risultati parzialmente corretti e vuoti irriducibili — mantiene la sua forza come principio euristico di straordinaria produttività. Esso identifica il pattern ricorrente nella storia intellettuale del XX e XXI secolo in cui la formalizzazione incontrava i propri limiti strutturali: in logica, in teoria dell'informazione, nella teoria delle decisioni, nella filosofia della mente. Il valore di questo principio non è di ordine strettamente deduttivo ma di ordine abduttivo: suggerisce dove cercare i limiti di nuovi sistemi formali prima che si

Silvano Salvador

manifestino empiricamente, e orienta la ricerca verso soluzioni che non pretendano di colmare i vuoti dall'interno del sistema ma che gestiscano l'ineliminabilità di quei vuoti attraverso architetture più flessibili.

III-ter. Le critiche esterne al paradigma ottimizzante: Marcus, Bender e la questione della comprensione

Il paradigma del reinforcement learning profondo e dell'allineamento basato su feedback umano non ha solo limiti interni — le cinque dimensioni di fragilità del RLHF, la convergenza strumentale, l'impossibilità di Arrow — ma è oggetto di critiche esterne che mettono in discussione alcune delle sue assunzioni fondamentali. Presentare queste critiche con accuratezza è un requisito di integrità intellettuale, anche quando esse non sono conclusive.

La prima linea critica riguarda la sistematicità e la composizionalità. Gary Marcus, in una serie di lavori dal 2018 al 2022, ha argomentato che i sistemi neurali profondi — inclusi i grandi modelli linguistici — apprendono correlazioni distribuzionali ma non strutture composizionali genuine nel senso di Fodor e Pylyshyn: non rappresentano relazioni tra simboli astratti in modo sistematico e separabile. Il benchmark SCAN di Lake e Baroni (2018) ha mostrato che i modelli seq2seq standard falliscono sistematicamente su combinazioni di istruzioni che non appaiono nel training set, nonostante le istruzioni siano composte da componenti già osservati. I grandi modelli linguistici contemporanei mostrano capacità composizionali emergenti che sfidano parzialmente questa diagnosi, ma la risposta standard — "con abbastanza dati e parametri, la composizionalità emerge" — evita piuttosto che risponde

Silvano Salvador

alla questione di se quella composizionalità sia sistematica nel senso richiesto da una teoria cognitiva.

La seconda linea critica riguarda la natura dei dati di addestramento e il rischio di amplificazione di bias. Bender, Gebru, McMillan-Major e Shmitchell (2021), nel loro controverso articolo "On the Dangers of Stochastic Parrots", hanno argomentato che i modelli linguistici addestrati su corpora di testo a larga scala non "comprendono" il linguaggio ma producono pattern statistici dall'archivio di testo umano, con tre rischi strutturali: l'amplificazione dei bias demografici e culturali presenti in quel corpus (il testo non è rappresentativo dell'umanità ma delle popolazioni con accesso alla scrittura digitale); il costo ambientale del training a larga scala; la dissociazione tra competenza linguistica e competenza nel mondo, che produce sistemi credibili ma non affidabili. Queste critiche sono parzialmente sovrapponibili ai limiti del RLHF discussi nel capitolo precedente — il problema della non-rappresentatività dei valutatori e quello delle allucinazioni come mismatch distribuzionale — ma le inquadrano non come problemi tecnici da risolvere ma come caratteristiche strutturali del paradigma.

La valutazione corretta di queste critiche richiede di distinguere tra ciò che esse dimostrano, ciò che suggeriscono e ciò che restano congetture. Ciò che è dimostrato empiricamente: i benchmark di composizionalità mostrano limiti sistematici nei modelli standard, e questi limiti migliorano ma non scompaiono con la scala. Ciò che è ben motivato: i bias demografici nei corpora si trasmettono nei modelli in modo misurabile, e il RLHF non li elimina ma li redistribuisce secondo le preferenze dei valutatori. Ciò che rimane congetturale: la tesi che i modelli "non comprendano davvero" dipende da una definizione di comprensione che la critica non fornisce in forma operativa e falsificabile; senza questa definizione, la critica rischia di essere

Silvano Salvador

empiricamente vuota. La risposta informazionale a questa obiezione è che la comprensione, nel senso operativo rilevante per l'allineamento, può essere approssimata come la capacità di agire correttamente in distribuzione e fuori distribuzione — e su questa definizione i modelli grandi mostrano progressi sostanziali, anche se non definitivi.

III-quater. La governance istituzionale dell'allineamento: strutture distribuite e meccanismi di controllo

Le tre direzioni del programma positivo discusse nella sezione precedente — feedback iterativo, incertezza sugli obiettivi, interpretabilità meccanicistica — sono soluzioni tecniche a un problema che ha anche una dimensione istituzionale irriducibile. La domanda "chi decide cosa un sistema di AI dovrebbe ottimizzare?" non ha risposta tecnica: è una domanda di design istituzionale che tocca la teoria democratica, la filosofia politica e l'economia del welfare.

Il framework teorico più sviluppato per la governance distribuita dell'allineamento deriva dalla teoria del mechanism design e dalla social choice theory. Dati n gruppi portatori di interessi S_1, ..., S_n con funzioni di utilità u_1, ..., u_n, un meccanismo di governance G : U^n → R mappa i profili di preferenze riportati su una funzione di ricompensa aggregata. Il teorema di Gibbard-Satterthwaite (1973) afferma che, per $|X| \geq 3$ alternative, non esiste meccanismo non-ditatoriale e strategy-proof: qualsiasi meccanismo non banale può essere manipolato da agenti strategici. Questo è il correlato, nel dominio dell'aggregazione istituzionale, del teorema di Arrow discusso nel capitolo quarto.

Silvano Salvador

La risposta operativa a questa impossibilità non è l'abbandono della governance ma il design di meccanismi che la gestiscano. Quattro architetture istituzionali sono state proposte e parzialmente implementate. La prima è la governance a più livelli (multi-stakeholder governance): i valori operativi del sistema di AI sono determinati da un consiglio che include sviluppatori, utenti finali, rappresentanti istituzionali e parti civili, con meccanismo di voto approssimativo e potere di veto su dimensioni di sicurezza. Questa architettura è attualmente adottata, in forme diverse, da Anthropic (Constitutional AI con principi multi-fonte), da DeepMind (safety board) e dalla regolamentazione europea (EU AI Act con obbligo di fundamental rights impact assessment per sistemi ad alto rischio).

La seconda architettura è l'audit e la certificazione di terze parti. Un'entità indipendente dal developer valuta il sistema rispetto a standard predefiniti di sicurezza e allineamento, analogamente alla certificazione medica o avionica. La differenza strutturale è che i sistemi di AI sono molto più difficili da testare in modo esaustivo dei dispositivi fisici, per le ragioni discusse nel capitolo sull'interpretabilità: la distribuzione degli input possibili è inaccessibile, il comportamento fuori distribuzione è imprevedibile, e le capacità emergenti possono manifestarsi a soglie di scala non anticipate.

La terza architettura è la trasparenza computazionale selettiva: rendere ispezionabili specifici aspetti del comportamento del sistema — le sue risposte su benchmark di sicurezza, la distribuzione statistica dei comportamenti su dataset rappresentativi, il comportamento su "red team" inputs — senza richiedere la divulgazione completa dei pesi, che creerebbe rischi di sicurezza propri. Questa architettura è compatibile con il programma di interpretabilità meccanicistica: man mano che i

Silvano Salvador

metodi per identificare circuiti e rappresentazioni migliorano, la trasparenza computazionale selettiva diventa più informativa.

La quarta architettura è la correggibilità regolamentare: sistemi di AI progettati per essere modificabili da autorità istituzionali designate, con meccanismi tecnici che permettano l'aggiornamento dei vincoli di comportamento in risposta a nuove informazioni o a cambiamenti normativi. Questa architettura corrisponde, a livello tecnico, al framework CIRL di Russell discusso nel capitolo quarto: l'agente mantiene incertezza sugli obiettivi e è costruito per accettare correzioni esterne come evidenza informativa sugli obiettivi veri piuttosto che come interferenze da resistere.

Nessuna di queste architetture risolve il problema dell'allineamento nei propri termini; ciascuna lo gestisce parzialmente. La combinazione delle quattro, operata a livello istituzionale con appropriati meccanismi di incentivo e responsabilità, costituisce il programma positivo realista: non la soluzione dell'impossibilità di Arrow, ma la costruzione di strutture istituzionali che riducano l'impatto della dittatura preferenziale e aumentino la probabilità che i sistemi di AI riflettano valori condivisi piuttosto che quelli del subset demografico che ha progettato o addestratto il sistema.

IV. La questione filosofica irrisolta: coscienza e comprensione

Il trio di saggi ha sistematicamente evitato di pronunciarsi sulla domanda filosofica più profonda che l'intelligenza artificiale pone: se i sistemi neurali profondi abbiano comprensione, coscienza, esperienza soggettiva. Non per incapacità di affrontarla, ma perché questa è la

Silvano Salvador

domanda che la matematica e la teoria dell'informazione non sono — ancora, forse mai — in grado di rispondere in modo conclusivo.

La disputa tra Searle e i suoi critici — la "stanza cinese" come argomento contro la sufficienza della sintassi per la semantica — non è risolvibile con gli strumenti che abbiamo illustrato. La domanda di Nagel — "c'è qualcosa che fa" essere un pipistrello, o un grande modello linguistico? — non ha una risposta nella teoria delle distribuzioni di probabilità. La questione di se la comprensione semantica richieda qualcosa di più della struttura distribuzionale del significato, intesa come pattern statistici appresi da grandi corpora, dipende da premesse metafisiche che la matematica non può stabilire.

Quello che la matematica può stabilire, come abbiamo cercato di stabilire con rigore, è la struttura computazionale dell'intelligenza: i suoi meccanismi di apprendimento, i suoi limiti formali, le proprietà emergenti che si producono dall'ottimizzazione su distribuzioni sufficientemente ricche. Questa struttura computazionale è necessaria per qualsiasi risposta adeguata alle domande filosofiche più profonde — non si può discutere in modo informato della coscienza dell'intelligenza artificiale senza capire cosa essa faccia computazionalmente. Ma non è sufficiente: le domande fenomenologiche rimangono aperte oltre il perimetro della matematica, e questa apertura è onestà intellettuale, non una lacuna da colmare.

V. Coda: Wittgenstein e il linguaggio dell'intelligenza

"Di ciò di cui non si può parlare, si deve tacere." L'ultima proposizione del Tractatus logico-philosophicus di Wittgenstein è spesso letta come un'ingiunzione al silenzio davanti ai limiti del linguaggio. Ma Wittgenstein

Silvano Salvador

stesso, nelle Ricerche filosofiche, ha capovolto questa visione sostituendo il modello del linguaggio come calcolo formale con quello del linguaggio come pratica — insieme di giochi linguistici radicati nelle forme di vita, dove il significato non è la corrispondenza con i fatti ma l'uso nelle pratiche condivise.

La storia dell'intelligenza artificiale è, in una certa misura, la storia di questa transizione giocata su scala computazionale. Il paradigma simbolico era il Tractatus: il linguaggio come calcolo formale, il significato come corrispondenza con strutture simboliche, l'intelligenza come deduzione da premesse esplicite. Il paradigma statistico è le Ricerche: il linguaggio come distribuzione di uso su un corpus vastissimo, il significato come struttura del contesto distributionale, l'intelligenza come approssimazione di un prior universale appresa dall'immersione nei dati del mondo.

Le tre impossibilità che abbiamo delineato delimitano ciò di cui si può parlare con rigore: i confini entro cui la ricerca può muoversi con strumenti formali precisi. Al di là di questi confini — nella fenomenologia dell'esperienza soggettiva, nella valutazione morale ultima degli obiettivi, nel giudizio su cosa rende una vita umana degna di essere vissuta — il linguaggio formale cede necessariamente il passo ad altri modi di comprensione. Ma questa cessione non è sconfitta: è la mappa che indica dove finisce un tipo di conoscenza e ne inizia un altro. Di ciò di cui si può parlare con precisione crescente, bisogna continuare a parlare — sapendo che la precisione assoluta è irraggiungibile, ma che la direzione dell'avvicinamento è chiara, e che questo avvicinamento è, esso stesso, il progresso.

Silvano Salvador

V-bis. Apprendimento federato e privacy differenziale: vincoli strutturali sull'addestramento

Il paradigma standard dell'addestramento di reti neurali profonde richiede che i dati di training siano centralizzati su un server: l'ottimizzazione stocastica del gradiente aggiorna i parametri del modello iterando sull'intero dataset. Questo requisito di centralizzazione è incompatibile con i vincoli legali (GDPR, HIPAA) e pratici (dati sensibili, dispositivi edge) di molte applicazioni reali. Il Federated Learning (McMahan et al., 2017) risolve questo conflitto invertendo la direzione del trasferimento: invece di spostare i dati verso il modello, si sposta il modello verso i dati.

L'algoritmo FedAvg aggiorna il modello globale attraverso una media pesata degli aggiornamenti locali. Ad ogni round t, il server invia i parametri correnti θ_t a un sottoinsieme di K client; ogni client k esegue E step di SGD sul proprio dataset locale D_k producendo θ_t^k; il server aggrega:

$$\theta_{t+1} = \sum_k \left(|D_k| \,/\, \sum_j |D_j|\right) \cdot \theta_t^k$$

Il problema fondamentale di FedAvg è la non-iid: i dataset locali dei diversi client non sono identicamente distribuiti, e questa eterogeneità causa client drift — i modelli locali ottimizzati su D_k divergono nella direzione che massimizza le performance locali, che non coincide necessariamente con la direzione di miglioramento del modello globale. FedProx (Li et al., 2020) mitiga il drift aggiungendo un termine di prossimità alla perdita locale che penalizza la distanza dal modello globale:

Silvano Salvador

$$\min_{\theta} h_k(\theta) + \mu/2 \cdot ||\theta - \theta_t||^2$$

La privacy differenziale (Dwork e Roth, 2014) formalizza la garanzia di privacy per i dati dei singoli client. Un algoritmo randomizzato M è (ε, δ)-differentially private se, per ogni coppia di dataset adiacenti D e D' che differiscono in un solo elemento, e per ogni sottoinsieme S di output:

$$P[M(D) \in S] \leq e^{\varepsilon} \cdot P[M(D') \in S] + \delta$$

In DP-SGD (Abadi et al., 2016), la garanzia differenziale è ottenuta attraverso due meccanismi: il clipping dei gradienti individuali ($||\nabla_i|| \leq C$) che limita la sensibilità del gradiente medio, e l'aggiunta di rumore gaussiano calibrato sulla sensibilità $N(0, \sigma^2C^2I)$. Il parametro ε controlla il trade-off privacy-utility: valori di ε piccoli (alta privacy) richiedono rumore elevato che degrada la performance del modello. Per i grandi modelli linguistici, raggiungere garanzie di privacy stringenti ($\varepsilon \leq 1$) con DP-SGD produce una degradazione significativa della qualità: il rumore necessario è proporzionale alla dimensionalità dello spazio dei gradienti, che è nell'ordine dei miliardi di parametri.

Per il problema dell'allineamento, l'apprendimento federato introduce una tensione strutturale: la personalizzazione del comportamento del modello ai contesti e ai valori dei singoli utenti — obiettivo desiderabile per l'allineamento fine-grained — richiede accesso ai dati locali che la privacy differenziale limita. Questa tensione non ha soluzione tecnica perfetta: è un caso particolare del trade-off informazionale fondamentale tra personalizzazione e privacy, formalizzabile come trade-off tra informazione mutua I(modello; dati_utente) e garanzia di indistinguibilità richiesta dalla privacy differenziale. Più il modello è personalizzato, più rivela informazioni sui dati dell'utente che lo hanno personalizzato.

Silvano Salvador

Appendice Tecnica I

Dimostrazioni complete e strutture matematiche avanzate

A.1 — Derivazione completa dell'equazione di Bellman e proprietà spettrali dell'operatore T*

In questa appendice si fornisce la derivazione rigorosa delle proprietà dell'operatore di ottimalità di Bellman T*, con attenzione alle condizioni di regolarità e alle strutture spettrali che determinano la velocità di convergenza dell'iterazione del valore.

Si lavora nello spazio delle funzioni limitate $\mathcal{B}(\mathcal{S}) = \{V : \mathcal{S} \to \mathbb{R} : \sup_{s} |V(s)| < \infty\}$ equipaggiato con la norma del sup $\|V\|_\infty = \sup_{s \in \mathcal{S}} |V(s)|$. Si è già mostrato che $(T^*V)(s) = \max_{a \in \mathcal{A}} [R(s,a) + \gamma \sum_{s'} P(s'|s,a) V(s')]$ è una contrazione con coefficiente γ. Vale la pena caratterizzarne ulteriormente le proprietà spettrali per comprendere la struttura della convergenza.

Sia V^* il punto fisso di T^*, e sia $V_k = T^{*k} V_0$ la sequenza delle iterate partendo da V_0. Si ha:

$$\|V_{k+1} - V_k\|_\infty = \|T^* V_k - T^* V_{k-1}\|_\infty \leq \gamma \|V_k - V_{k-1}\|_\infty \leq \gamma^k \|V_1 - V_0\|_\infty$$

e per la disuguaglianza triangolare:

$$\|V_k - V^*\|_\infty \leq \sum_{j=k}^{\infty} \|V_{j+1} - V_j\|_\infty \leq \gamma^k \|V_1 - V_0\|_\infty / (1 - \gamma)$$

Silvano Salvador

Questo è il bound a priori sulla convergenza. Esiste anche un bound a posteriori più stretto, basato sul residuo di Bellman $r_k = \|T^* V_k - V_k\|_\infty$:

$$\|V_k - V^*\|_\infty \leq r_k / (1 - \gamma)$$

La dimostrazione usa la stima: $V^* = T^* V^*$ e quindi $\|V^* - V_k\|_\infty = \|T^* V^* - V_k\|_\infty \leq \|T^* V^* - T^* V_k\|_\infty + \|T^* V_k - V_k\|_\infty \leq \gamma \|V^* - V_k\|_\infty + r_k$, *da cui* $\|V^* - V_k\|_\infty \leq r_k / (1 - \gamma)$. Questo bound a posteriori è utile in pratica perché può essere calcolato senza conoscere V^* — richiede soltanto di calcolare il residuo di Bellman della stima corrente.

A.2 — La struttura della dualità nel minimax: connessione con i saddle point

Il teorema minimax di von Neumann può essere visto come un caso speciale di un risultato molto più generale sui saddle point di funzioni bilineari su prodotti di insiemi compatti convessi. Sia $f : X \times Y \to \mathbb{R}$ una funzione continua con X, Y compatti convessi, tale che $f(x, \cdot)$ è convessa per ogni x fissato e $f(\cdot, y)$ è concava per ogni y fissato. Allora:

$$\max_{x \in X} \min_{y \in Y} f(x, y) = \min_{y \in Y} \max_{x \in X} f(x, y)$$

Nel caso del teorema minimax, $f(p, q) = p^T M q$ è bilineare in (p, q) — convessa e concava, quindi convex-concave — e $X = \Delta_m$, $Y = \Delta_n$ sono simplessi compatti e convessi. Il teorema del minimax segue dal teorema di saddle point di Von Neumann, che a sua volta segue dal teorema di Brouwer o da tecniche di dualità LP come mostrato nel testo principale.

Silvano Salvador

Il punto di sella (p^*, q^*) è caratterizzato da:

$$\max_{p} \; p^{T} M q^{*} = p^{T} M q = \min_{q} \; p^{*T} M q$$

Le condizioni di Karush-Kuhn-Tucker per il programma lineare (P_1) sono:

$$\sum_i M_{ij} \, p_i \geq V \;\forall j, \; p_j \geq 0, \; \sum_j p_j \left(\sum_i M_{ij} \, p_i - V\right) = 0$$

Le condizioni di complementarità $p_j > 0 \implies \sum_i M_{ij} \, p_i = V^*$ affermano che la strategia mista ottimale di P_1 deve assegnare probabilità positiva soltanto alle azioni pure che garantiscono esattamente il valore V^* contro la strategia ottimale q^* di P_2 — le azioni per cui P_2 è "indifferente". Simmetricamente, la strategia ottimale q^* assegna probabilità positiva soltanto alle azioni pure per cui P_1 riceve esattamente V^*. Questo è il principio dell'indifferenza (indifference principle) della teoria dei giochi a somma zero: in un equilibrio in strategie miste, ogni giocatore è indifferente tra le azioni nel supporto della propria strategia ottimale.

A.3 — Dimostrazione completa del teorema di Arrow

Si dimostra qui il teorema di impossibilità di Arrow con la massima precisione formale, seguendo la presentazione di Sen (1979) con alcune semplificazioni espositive.

Sia $N = \{1, \ldots, n\}$ l'insieme degli individui ($n \geq 2$), sia X l'insieme delle alternative con $|X| \geq 3$, e sia $\mathscr{P}$ lo spazio degli ordinamenti completi e transitivi su X (relazioni di preferenza razionali). Un profilo di preferenze è un elemento $R = (R_1, \ldots, R_n) \in \mathscr{P}^n$. Una funzione di scelta sociale

Silvano Salvador

(SWF) è una funzione $F : \mathscr{P}^n \to \mathscr{P}$ che associa a ogni profilo di preferenze un ordinamento collettivo.

Le condizioni di Arrow sono:

U (Dominio universale): $Dom(F) = \mathscr{P}^n$. P (Pareto debole): Per ogni $R \in \mathscr{P}^n$ e ogni coppia $x, y \in X$, se xR_iy per ogni $i \in N$ (unanimità stretta: tutti preferiscono strettamente x a y), allora $x\ F(R)\ y$. IIA (Indipendenza dalle alternative irrilevanti): Per ogni $R, R' \in \mathscr{P}^n$ e ogni coppia $x, y \in X$, se per ogni $i \in N$ i confronti di R_i e R'_i su $\{x, y\}$ coincidono ($xR_iy \Leftrightarrow xR'_iy$ e $yR_ix \Leftrightarrow yR'_ix$), allora i confronti di F(R) e F(R') su $\{x, y\}$ coincidono. ND (Non-dittatorialità): Non esiste $i \in N$ tale che per ogni $R \in \mathscr{P}^n$ e ogni $x, y \in X$, se xR_iy allora $x\ F(R)\ y$. T (Transitività): $F(R) \in \mathscr{P}$ per ogni R — l'ordinamento collettivo è transitivo.

Teorema (Arrow, 1951). Non esiste F che soddisfi simultaneamente U, P, IIA, ND e T con $|X| \geq 3$.

Definizione. Un individuo i è decisivo sulla coppia (x, y) per il profilo R se xR_iy implica $x\ F(R)\ y$, indipendentemente dalle preferenze degli altri. Una coalizione $C \subseteq N$ è decisiva su (x, y) se, per ogni profilo in cui tutti i membri di C preferiscono x a y, la SWF preferisce x a y.

Lemma 1 (Field expansion). Se una coalizione C è quasi-decisiva su una coppia (x, y) — ovvero decisiva in almeno un profilo — allora è decisiva su tutte le coppie.

Dimostrazione del Lemma 1. Si supponga C quasi-decisiva su (x, y): esiste un profilo R_0 in cui ogni $i \in C$ ha $x\ R_{0_i}\ y$ e $x\ F(R_0)\ y$, qualunque siano le preferenze degli altri su x e y. Mostriamo che C è decisiva su una coppia generica (a, b) con $a \neq b$.

Caso (i): $b \neq x$, $a \neq y$ (nessuna sovrapposizione). Per U, si costruisce un profilo R_1 in cui ogni $i \in C$ ha $a\ R_{1_i}\ x\ R_{1_i}\ y\ R_{1_i}\ b$, e ogni $j \notin C$ ha y

Silvano Salvador

R_1_j b R_1_j a R_1_j x. Per Pareto (applicato a y R_1_i b per tutti i), y F(R_1) b. Per la quasi-decisività di C su (x, y) applicata alle posizioni relative di x e y in R_1, x F(R_1) y. Per transitività: x F(R_1) b. Le preferenze su (a, x) di C sono a R_1_i x per tutti i, e le preferenze degli altri su (a, x) sono arbitrarie (irrilevanti per IIA). Per Pareto su (a, x) (poiché tutti preferiscono a a x), a F(R_1) x. Per transitività: a F(R_1) b. Per IIA (le preferenze di tutti su (a, b) nel profilo R_1 sono a $\succ$ b per $i \in C$, qualsiasi per $j \notin C$), C è decisiva su (a, b) in generale.

I casi rimanenti si trattano in modo analogo con appropriate permutazioni delle preferenze. □

Lemma 2 (Group contraction). Se una coalizione C con $|C| \geq 2$ è decisiva, allora esiste una sottocolalizione propria $C' \subsetneq C$ che è decisiva.

Dimostrazione del Lemma 2. Siano $i, j \in C$ con $i \neq j$, e siano $C_1 = \{i\}$ e $C_2 = C \setminus \{i\}$. Per U, si costruisce un profilo R_2 tale che: i ha x R_2_i y R_2_i z, ogni $k \in C_2$ ha y R_2_k z R_2_k x, ogni $j \notin C$ ha z R_2_j x R_2_j y. Per la decisività di C su (x, y) (i $\in$ C preferisce x a y, $C_2 \subseteq$ C preferisce y a x — ma il profilo R_2 è in C-decisivo per x contro y soltanto se tutti in C preferiscono x a y).

Si usa invece la seguente costruzione: il profilo in cui tutti in C preferiscono x a y produce x F(R) y per decisività di C. In questo profilo, il sottoinsieme $C_1 = \{i\}$ ha x R y, e $C_2 = C\{i\}$ ha y R z R x. Si argomenta per casi: o $C_1 = \{i\}$ è quasi-decisivo su (x, z), o C_2 è quasi-decisivo su (y, z). Per il Lemma 1, la coalizione quasi-decisiva diventa decisiva su tutte le coppie. □

Dimostrazione del teorema di Arrow. La coalizione N di tutti gli individui è decisiva per Pareto (P). Per applicazione iterata del Lemma 2, esiste un singolo individuo $\{d\} \subseteq$ N che è decisivo su tutte le coppie.

Silvano Salvador

Ma un individuo decisivo su tutte le coppie è per definizione un dittatore, contraddicendo ND. □

A.4 — Derivazione rigorosa del gradiente della perdita di Bradley-Terry e convergenza del reward model

La perdita di Bradley-Terry $L_{BT}(\varphi) = -\mathbb{E}_{(x, y_w, y_l) \sim D} [\log \sigma(R_\varphi(x, y_w) - R_\varphi(x, y_l))]$ è convessa in $R_\varphi(x, y_w) - R_\varphi(x, y_l)$ per ogni coppia fissata, come si verifica dal segno della derivata seconda:

$d^2/d\Delta^2 [-\log \sigma(\Delta)] = \sigma(\Delta)(1 - \sigma(\Delta)) > 0$

Il gradiente rispetto ai parametri φ del reward model è:

$\nabla_\varphi L_{BT}(\varphi) = -\mathbb{E}_{D} [(1 - \sigma(R_\varphi(x, y_w) - R_\varphi(x, y_l))) \nabla_\varphi R_\varphi(x, y_w) - \sigma(R_\varphi(x, y_w) - R_\varphi(x, y_l)) \nabla_\varphi R_\varphi(x, y_l) \cdot (-1)]$

$= -\mathbb{E}_{D} [\sigma(R_\varphi(x, y_l) - R_\varphi(x, y_w)) (\nabla_\varphi R_\varphi(x, y_w) - \nabla_\varphi R_\varphi(x, y_l))]$

Il fattore $\sigma(R_\varphi(x, y_l) - R_\varphi(x, y_w))$ è la probabilità di classificazione errata: è grande (vicino a 1/2) quando la differenza di reward tra y_w e y_l è vicina a zero (il modello è incerto), e piccolo (vicino a 0) quando la differenza è grande e positiva (il modello distingue correttamente con alta confidenza). Il gradiente è quindi concentrato sulle comparazioni difficili, con un meccanismo che assomiglia all'AdaBoost ma nella direzione opposta: invece di concentrare il peso sugli esempi classificati erroneamente, il gradient descent naturalmente si focalizza sugli esempi difficili attraverso la struttura della funzione logistica.

La convergenza del training del reward model a un minimo locale è garantita dalla convessità locale di L_{BT} in $R_\varphi(x, y_w) - R_\varphi(x, y_l)$

Silvano Salvador

per ogni coppia fissata. La convessità globale in φ richiede che R_φ sia lineare in φ — il che è vero per le reti neurali soltanto nell'ultimo strato, non in generale. Per reti neurali profonde, L_{BT}(φ) è in generale non convessa in φ, e si garantisce soltanto la convergenza a punti stazionari critici.

A.5 — La struttura dell'overoptimization: il modello di Gao-Ngo-Lowe-Clark

Gao, Ngo, Lowe e Clark (2023) hanno sviluppato un modello teorico della curva di overoptimization che cattura matematicamente il deterioramento della qualità reale al crescere dell'intensità dell'ottimizzazione del reward model proxy. Sia R_proxy il reward model addestrato su N comparazioni umane, e sia R_gold un reward model "di riferimento" addestrato su un dataset molto più grande (o, idealmente, le preferenze umane reali). Si assume che entrambi siano realizzazioni rumorose della stessa funzione di ricompensa reale U:

R_proxy(y) = U(y) + ε_proxy(y), ε_proxy ~ N(0, σ²_proxy) R_gold(y) = U(y) + ε_gold(y), ε_gold ~ N(0, σ²_gold)

con σ²_gold ≪ σ²_proxy. Se si ottimizza la politica rispetto a R_proxy usando PPO e si usa R_gold come misura della qualità reale, il comportamento atteso è il seguente. All'inizio dell'ottimizzazione, R_proxy e R_gold sono correlati positivamente: migliorare R_proxy tende a migliorare R_gold. Man mano che l'ottimizzazione progredisce, la politica trova risposte che massimizzano ε_proxy in modo sistematico — risposte con alta varianza di stima positiva nel reward model proxy — e questi "buchi" del proxy non corrispondono a miglioramenti reali. La qualità R_gold cresce inizialmente, raggiunge un picco al valore K* =

Silvano Salvador

$\sqrt{(2\ \sigma^2_proxy\ /\ \sigma^2_gold)} \cdot \log(N)$ (in unità appropriate), poi decresce monotonicamente.

Formalmente, il ritorno atteso di R_gold dopo K passi PPO è approssimativamente:

$$\mathbb{E}[R_gold(\pi_K)] \approx \mu_U + \sqrt{(2\ \sigma^2_gold\ /\ \sigma^2_proxy)} \cdot \sqrt{K} - K \cdot \sigma^2_proxy\ /\ (2\ \mu_U)$$

dove il primo termine $\sqrt{K}$ cresce per l'allineamento genuino e il secondo termine $-K$ cresce per il reward hacking. Il massimo si trova a $K^* = \mu_U\ \sqrt{(\sigma^2_gold\ /\ \sigma^2_proxy)}\ /\ \sigma^2_proxy$, e il picco di qualità è proporzionale a $\sqrt{(\sigma^2_gold\ /\ \sigma^2_proxy)}$ — la radice quadrata del rapporto tra la varianza del proxy e quella del gold. Questo modello afferma che la robustezza del RLHF all'overoptimization è proporzionale alla qualità del reward model, misurata dal rapporto $\sigma^2_gold\ /\ \sigma^2_proxy$: migliore è il reward model (piccolo σ^2_proxy relativo a σ^2_gold), più a lungo si può ottimizzare senza degrado.

Nota bibliografica essenziale

I risultati tecnici citati in questo saggio rimandano alle fonti primarie seguenti, raggruppate per tema. Per la teoria degli MDP e la programmazione dinamica: Bellman (1957), Dynamic Programming; Puterman (1994), Markov Decision Processes; Bertsekas e Tsitsiklis (1996), Neuro-Dynamic Programming. Per il Q-learning e le DQN: Watkins e Dayan (1992), Q-learning; Mnih e colleghi (2015), Human-Level Control through Deep Reinforcement Learning; van Hasselt, Guez e Silver (2016), Deep Reinforcement Learning with Double Q-learning; Wang e colleghi (2016), Dueling Network

Silvano Salvador

Architectures. Per i policy gradient e le varianti: Williams (1992), Simple Statistical Gradient-Following Algorithms; Sutton e colleghi (2000), Policy Gradient Methods for Reinforcement Learning; Kakade (2002), A Natural Policy Gradient; Schulman e colleghi (2015), Trust Region Policy Optimization; Schulman e colleghi (2016), High-Dimensional Continuous Control Using Generalized Advantage Estimation; Schulman e colleghi (2017), Proximal Policy Optimization Algorithms. Per la connessione neuroscientifica: Schultz, Dayan e Montague (1997), A Neural Substrate of Prediction and Reward. Per la teoria dei giochi: von Neumann e Morgenstern (1944), Theory of Games and Economic Behavior; Nash (1950), Equilibrium Points in N-Person Games; Taylor e Jonker (1978), Evolutionary Stable Strategies and Game Dynamics; Fudenberg e Maskin (1986), The Folk Theorem in Repeated Games with Discounting or with Incomplete Information. Per AlphaZero: Silver e colleghi (2018), A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go. Per l'allineamento e Goodhart: Arrow (1951), Social Choice and Individual Values; Goodhart (1975), Problems of Monetary Management; Christiano e colleghi (2017), Deep Reinforcement Learning from Human Preferences; Ouyang e colleghi (2022), Training Language Models to Follow Instructions with Human Feedback; Gao e colleghi (2023), Scaling Laws for Reward Model Overoptimization; Krakovna e colleghi (2020), Specification Gaming: the Flip Side of AI Ingenuity. Per l'interpretabilità meccanicistica: Olah e colleghi (2020), Zoom In: An Introduction to Circuits; Elhage e colleghi (2022), Toy Models of Superposition; Meng e colleghi (2022), Locating and Editing Factual Associations in GPT; Cunningham e colleghi (2023), Sparse Autoencoders Find Highly Interpretable Features in Language Models. Per AIXI: Hutter (2005), Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability.

Silvano Salvador

Appendice Tecnica II

Policy Gradient, gradiente naturale e TRPO: derivazioni complete

A.II.1 — Preliminari: lo spazio delle traiettorie e la struttura di probabilità

Prima di avviare la derivazione del teorema del policy gradient è opportuno fissare con cura la struttura degli oggetti matematici in gioco, perché una gran parte delle difficoltà tecniche del campo deriva dall'imprecisione in questo passaggio preliminare.

Sia dato un MDP $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ con politica differenziabile $\pi_\theta : \mathcal{S} \to \Delta(\mathcal{A})$ parametrizzata da $\theta \in \mathbb{R}^d$. Una traiettoria di lunghezza T è una sequenza $\tau = (s_0, a_0, s_1, a_1, \ldots, s_{T-1}, a_{T-1}, s_T)$. Lo spazio delle traiettorie $\mathcal{T}$ è il prodotto cartesiano $(\mathcal{S} \times \mathcal{A})^T \times \mathcal{S}$. La misura di probabilità su $\mathcal{T}$ indotta dalla politica π_θ e dalle dinamiche P è:

$$p_\theta(\tau) = d_0(s_0) \cdot \prod_{t=0}^{T-1} \pi_\theta(a_t \mid s_t) \cdot P(s_{t+1} \mid s_t, a_t)$$

dove $d_0 : \mathcal{S} \to [0,1]$ è la distribuzione degli stati iniziali. Si noti che $p_\theta(\tau)$ dipende da θ esclusivamente attraverso i fattori $\pi_\theta(a_t \mid s_t)$; le dinamiche P e la distribuzione d_0 sono indipendenti da θ.

Il ritorno di una traiettoria è la somma scontata delle ricompense:

$$R(\tau) = \sum_{t=0}^{T-1} \gamma^t r_t, \quad r_t = R(s_t, a_t)$$

Silvano Salvador

L'obiettivo di performance è il ritorno atteso:

$$J(\theta) = \mathbb{E}_{\tau \sim p_\theta}[R(\tau)] = \int_{\mathcal{T}} p_\theta(\tau)\, R(\tau)\, d\tau$$

Si assume che $\pi_\theta(a \mid s)$ sia differenziabile rispetto a θ per ogni $(s, a) \in \mathcal{S} \times \mathcal{A}$, che le derivate parziali siano limitate uniformemente, e che lo scambio di integrazione e differenziazione sia lecito — condizioni soddisfatte per le parametrizzazioni standard come softmax, gaussiana e beta.

A.II.2 — Il log-derivative trick e la score function

Il trucco fondamentale su cui si fonda il teorema del policy gradient è l'identità:

$$\nabla_\theta p_\theta(\tau) = p_\theta(\tau) \cdot \nabla_\theta \log p_\theta(\tau)$$

la quale segue semplicemente dalla regola della catena applicata alla composizione logaritmo-esponenziale:

$$\nabla_\theta \log p_\theta(\tau) = \nabla_\theta p_\theta(\tau) \,/\, p_\theta(\tau) \Rightarrow \nabla_\theta p_\theta(\tau) = p_\theta(\tau) \cdot \nabla_\theta \log p_\theta(\tau)$$

Questa identità — nota come log-derivative trick o REINFORCE trick — permette di trasformare il gradiente di un'aspettativa in un'aspettativa di un gradiente, con la distribuzione di campionamento intatta:

$$\nabla_\theta J(\theta) = \nabla_\theta \int p_\theta(\tau)\, R(\tau)\, d\tau = \int \nabla_\theta p_\theta(\tau) \cdot R(\tau)\, d\tau = \int p_\theta(\tau) \cdot \nabla_\theta \log p_\theta(\tau) \cdot R(\tau)\, d\tau = \mathbb{E}_{\tau \sim p_\theta}[\nabla_\theta \log p_\theta(\tau) \cdot R(\tau)]$$

La quantità $\nabla_\theta \log p_\theta(\tau)$ è la funzione score della distribuzione delle traiettorie: essa misura, per ogni traiettoria τ, la sensibilità della log-probabilità di τ rispetto ai parametri θ. Ora si calcola esplicitamente:

Silvano Salvador

$\log p_\theta(\tau) = \log d_0(s_0) + \sum_{t=0}^{T-1} [\log \pi_\theta(a_t \mid s_t) + \log P(s_{t+1} \mid s_t, a_t)]$

Differenziando rispetto a θ:

$\nabla_\theta \log p_\theta(\tau) = \nabla_\theta \log d_0(s_0) + \sum_{t=0}^{T-1} [\nabla_\theta \log \pi_\theta(a_t \mid s_t) + \nabla_\theta \log P(s_{t+1} \mid s_t, a_t)] = 0 + \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) + 0 = \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t)$

I termini $\nabla_\theta \log d_0(s_0)$ e $\nabla_\theta \log P(s_{t+1} \mid s_t, a_t)$ sono identicamente zero perché né la distribuzione iniziale né le dinamiche dell'ambiente dipendono da θ. Si ottiene quindi:

$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta} [\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot R(\tau)]$

Questa è la formula REINFORCE grezza, ancora non ridotta nella varianza. Il passaggio cruciale per ricavare la forma standard del teorema del policy gradient è la riduzione della causalità.

A.II.3 — Riduzione per causalità e derivazione del teorema del policy gradient

L'argomento causale è il seguente. Il gradiente $\nabla_\theta \log \pi_\theta(a_t \mid s_t)$ *dipende dall'azione al tempo t. La ricompensa* $r_{t'}$ con $t' < t$ è già avvenuta al tempo t e non può essere influenzata dall'azione a_t: l'azione al passo t può influenzare soltanto le ricompense ai passi $t' \geq t$. Formalmente, sotto l'aspettativa rispetto alla politica π_θ, il termine $\nabla_\theta \log \pi_\theta(a_t \mid s_t)$ *è condizionalmente indipendente da* $r_{t'}$ dato s_t, per ogni $t' < t$. Quindi:

$\mathbb{E}_{\tau \sim p_\theta} [\nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot r_{t'}] = 0$ per ogni $t' < t$

Questo si dimostra decomponendo l'aspettativa:

Silvano Salvador

$\mathbb{E}_\{\tau\}\ [\nabla\theta\ log\ \pi_\theta(a_t\ |\ s_t)\ \cdot\ r\{t'\}] = \sum\{s_t\}\ \varrho\hat{}\{\pi_\theta\}(s_t)\ \sum\{a_t\}\ \pi_\theta(a_t\ |\ s_t)\ \nabla\theta\ log\ \pi_\theta(a_t\ |\ s_t) \cdot \mathbb{E}[r\{t'\}\ |\ s_t, a_t]$

dove $\varrho\hat{}\{\pi_\theta\}(s_t)$ è la distribuzione marginale sugli stati al tempo t, e si è usata la Markov property per fattorizzare. Ora, per $t' < t$, la ricompensa $r_\{t'\}$ è determinata da $(s_\{t'\}, a_\{t'\})$ con $t' < t$ e quindi è indipendente da a_t condizionalmente su s_t:

$\mathbb{E}[r_\{t'\}\ |\ s_t, a_t] = \mathbb{E}[r_\{t'\}\ |\ s_t]$ per $t' < t$

e pertanto:

$\sum\{a_t\}\ \pi_\theta(a_t\ |\ s_t)\ \nabla\theta\ log\ \pi_\theta(a_t\ |\ s_t) \cdot \mathbb{E}[r\{t'\}\ |\ s_t] = \mathbb{E}[r\{t'\}\ |\ s_t] \cdot \sum\{a_t\}\ \pi_\theta(a_t\ |\ s_t) \cdot \nabla\theta\ log\ \pi_\theta(a_t\ |\ s_t) = \mathbb{E}[r\{t'\}\ |\ s_t] \cdot \sum\{a_t\}\ \nabla\theta\ \pi_\theta(a_t\ |\ s_t) = \mathbb{E}[r\{t'\}\ |\ s_t] \cdot \nabla\theta \sum\{a_t\}\ \pi_\theta(a_t\ |\ s_t) = \mathbb{E}[r_\{t'\}\ |\ s_t] \cdot \nabla_\theta\ 1 = 0$ □

Usando questo risultato, si può sostituire $R(\tau) = \sum\{t'=0\}\hat{}\{T-1\}\ \gamma\hat{}\{t'\}\ r\{t'\}$ con soltanto le ricompense future a partire da t, eliminando i termini che si annullano:

$\nabla\theta\ J(\theta)\ =\ \mathbb{E}\{\tau\ \sim\ p_\theta\}\ [\sum\{t=0\}\hat{}\{T-1\}\ \nabla\theta\ log\ \pi_\theta(a_t\ |\ s_t)\ \cdot\ \sum\{t'=t\}\hat{}\{T-1\}\ \gamma\hat{}\{t'\}\ r\{t'\}]$

Riscalando l'indice del ritorno futuro da t a 0 tramite l'estrazione di γ^t:

$= \mathbb{E}_\{\tau \sim p_\theta\}\ [\sum\{t=0\}\hat{}\{T-1\}\ \gamma^t\ \nabla\theta\ log\ \pi_\theta(a_t\ |\ s_t) \cdot \sum\{t'=t\}\hat{}\{T-1\}\ \gamma\hat{}\{t'-t\}\ r\{t'\}]$

Il termine $G_t = \sum\{t'=t\}\hat{}\{T-1\}\ \gamma\hat{}\{t'-t\}\ r\{t'\}$ è il ritorno futuro (return-to-go) a partire dal passo t. In aspettativa, condizionalmente sullo stato s_t e sull'azione a_t:

$\mathbb{E}[G_t\ |\ s_t, a_t] = Q\hat{}\{\pi_\theta\}(s_t, a_t)$

Si ottiene così il teorema del policy gradient nella sua forma standard:

Silvano Salvador

$\nabla \theta\ J(\theta) = \mathbb{E}\{\tau \sim \pi_\theta\}\ [\sum_\{t=0\}\^\{T-1\}\ \gamma^t\ \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot Q\^\{\pi_\theta\}(s_t, a_t)]$

Raccogliendo l'aspettativa sulla distribuzione degli stati $d\^\{\pi_\theta\}(s) = \sum_t \gamma^t P(s_t = s \mid \pi_\theta)$ (la misura di stato scontata), si scrive la forma compatta:

$\nabla \theta\ J(\theta) = \mathbb{E}\{s \sim d\^\{\pi_\theta\},\ a \sim \pi_\theta(\cdot \mid s)\}\ [\nabla_\theta \log \pi_\theta(a \mid s) \cdot Q\^\{\pi_\theta\}(s, a)]$ □

A.II.4 — Annullamento della baseline: dimostrazione formale

Si dimostra che sottrarre una baseline $b(s_t)$ dallo stimatore del gradiente non introduce distorsione. Per ogni funzione $b : \mathcal{S} \to \mathbb{R}$ e ogni passo t:

$\mathbb{E}_\{\tau \sim \pi_\theta\}\ [\nabla \theta\ log\ \pi_\theta(a_t \mid s_t) \cdot b(s_t)] = \mathbb{E}\{s_t \sim d\^\{\pi_\theta\}\}\ [b(s_t) \cdot \mathbb{E}_\{a_t \sim \pi_\theta(\cdot \mid s_t)\}\ [\nabla_\theta \log \pi_\theta(a_t \mid s_t)]]$

Ora si calcola il termine interno. Per la proprietà di normalizzazione della politica $\sum_\{a\}\ \pi_\theta(a \mid s) = 1$, differenziando rispetto a θ:

$\nabla \theta \sum\{a\}\ \pi_\theta(a \mid s) = \sum_\{a\}\ \nabla_\theta\ \pi_\theta(a \mid s) = 0$

Usando il log-derivative trick al contrario:

$\sum_\{a\}\ \nabla \theta\ \pi_\theta(a \mid s) = \sum\{a\}\ \pi_\theta(a \mid s)\ \nabla \theta\ log\ \pi_\theta(a \mid s) = \mathbb{E}\{a \sim \pi_\theta(\cdot \mid s)\}\ [\nabla_\theta \log \pi_\theta(a \mid s)] = 0$

Il termine interno è identicamente zero per ogni stato s. Quindi:

$\mathbb{E}_\{\tau \sim \pi_\theta\}\ [\nabla \theta\ log\ \pi_\theta(a_t \mid s_t) \cdot b(s_t)] = \mathbb{E}\{s_t\}\ [b(s_t) \cdot 0] = 0$ □

La baseline ottimale — quella che minimizza la varianza componente‑per‑componente dello stimatore del gradiente — si calcola

Silvano Salvador

come segue. Sia $\hat{g}^{\theta}_i$ la componente i-esima dello stimatore del gradiente per la traiettoria τ:

$$\hat{g}^{\theta}_i(\tau) = \sum_t \gamma^t \left[\partial_i \log \pi_\theta(a_t \mid s_t)\right] \cdot (G_t - b(s_t))$$

La varianza di $\hat{g}^{\theta}_i$ è:

$$\mathrm{Var}[\hat{g}^{\theta}_i] = \mathbb{E}[(\hat{g}^{\theta}_i)^2] - (\mathbb{E}[\hat{g}^{\theta}_i])^2$$

Minimizzando rispetto a b(s) con calcolo del gradiente e ponendolo a zero:

$$b^*_i(s) = \mathbb{E}[\sum_t (\partial_i \log \pi_\theta(a_t|s_t))^2 \, G_t \cdot \mathbf{1}[s_t = s]] \,/\, \mathbb{E}[\sum_t (\partial_i \log \pi_\theta(a_t|s_t))^2 \cdot \mathbf{1}[s_t = s]]$$

La baseline ottimale dipende dalla componente i, il che la rende computazionalmente infattibile nella pratica. L'approssimazione comune $b(s) = V^{\pi_\theta}(s)$ non è ottimale per nessuna singola componente, ma produce riduzioni di varianza significative e uniformi su tutte le componenti, essendo la stima imparziale del ritorno futuro medio.

A.II.5 — La funzione di vantaggio e la struttura actor-critic

Con la scelta $b(s_t) = V^{\pi_\theta}(s_t)$, il gradiente stimato diventa:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_t \gamma^t \nabla_\theta \log \pi_\theta(a_t \mid s_t) \cdot A^{\pi_\theta}(s_t, a_t)\right]$$

dove $A^{\pi_\theta}(s, a) = Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s)$ è la funzione di vantaggio. Si osservi che A^{π_θ} soddisfa la proprietà di media nulla:

$$\mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[A^{\pi_\theta}(s, a)] = \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[Q^{\pi_\theta}(s, a)] - V^{\pi_\theta}(s) = V^{\pi_\theta}(s) - V^{\pi_\theta}(s) = 0$$

Silvano Salvador

Questo chiarisce il significato del vantaggio: $A^{\pi_\theta}(s, a) > 0$ significa che l'azione a è migliore della media delle azioni in s sotto la politica π_θ; $A^{\pi_\theta}(s, a) < 0$ significa che è peggiore. Il gradiente del policy gradient spinge θ a aumentare la probabilità delle azioni con vantaggio positivo e a diminuire quella delle azioni con vantaggio negativo, proporzionalmente alla magnitudine del vantaggio.

In pratica, il vantaggio viene stimato attraverso un critic parametrico $V_\varphi(s)$ addestrato a minimizzare:

$$L_{\text{critic}}(\varphi) = \mathbb{E}_{\tau \sim \pi_\theta} [\sum_t (V_\varphi(s_t) - G_t)^2]$$

e la stima del vantaggio è $\hat{A}_t = G_t - V_\varphi(s_t)$. Il gradiente dell'actor è:

$$\nabla \theta J(\theta) \approx (1/N) \sum_{i=1}^{N} \sum_t \gamma^t \nabla_\theta \log \pi_\theta(a_t^{(i)} \mid s_t^{(i)}) \cdot \hat{A}_t^{(i)}$$

Il co-addestramento di actor e critic crea una struttura di bootstrap: il critic fornisce stime di valore che riducono la varianza del gradiente dell'actor; l'actor genera traiettorie che permettono al critic di aggiornarsi su nuovi stati. Questa interdipendenza richiede particolare attenzione alle frequenze di aggiornamento relativo: se il critic viene aggiornato troppo lentamente, le stime di vantaggio sono inaccurate; se viene aggiornato troppo velocemente, il gradiente dell'actor diventa instabile.

A.II.6 — Stima del vantaggio generalizzata (GAE): derivazione da prima

Si vuole stimare $A^{\pi_\theta}(s_t, a_t) = Q^{\pi_\theta}(s_t, a_t) - V^{\pi_\theta}(s_t)$. Si considerano le stime k-step del vantaggio, ottenute usando k passi di ricompense reali prima di stimare con il critic:

Silvano Salvador

$\hat{A}_t^{(k)} = -V_\varphi(s_t) + r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots + \gamma^{k-1} r_{t+k-1} + \gamma^k V_\varphi(s_{t+k})$

Questa stima è non distorta quando $V_\varphi = V^{\pi_\theta}$ esattamente, e tende al ritorno Monte Carlo per $k \to \infty$. Definendo il TD error a un passo come $\delta_t = r_t + \gamma V_\varphi(s_{t+1}) - V_\varphi(s_t)$, si verifica che:

$\hat{A}_t^{(k)} = \sum_{l=0}^{k-1} \gamma^l \delta_{t+l}$

Dimostrazione per induzione su k. Per $k = 1$: $\hat{A}_t^{(1)} = r_t + \gamma V_\varphi(s_{t+1}) - V_\varphi(s_t) = \delta_t$. ✓

Per il passo induttivo, assumendo $\hat{A}_t^{(k)} = \sum_{l=0}^{k-1} \gamma^l \delta_{t+l}$:

$\hat{A}_t^{(k+1)} = -V_\varphi(s_t) + r_t + \gamma (-V_\varphi(s_{t+1}) + r_{t+1} + \ldots + \gamma^{k-1} r_{t+k} + \gamma^k V_\varphi(s_{t+k+1})) = \delta_t + \gamma V_\varphi(s_{t+1}) - \gamma V_\varphi(s_{t+1}) + \gamma (-V_\varphi(s_{t+1}) + r_{t+1} + \ldots + \gamma^k V_\varphi(s_{t+k+1}))$

Raccogliendo: $\hat{A}_t^{(k+1)} = \delta_t + \gamma \hat{A}_{t+1}^{(k)} = \delta_t + \gamma \sum_{l=0}^{k-1} \gamma^l \delta_{t+1+l} = \sum_{l=0}^{k} \gamma^l \delta_{t+l}$. □

La stima GAE(γ,λ) è la media geometrica pesata di tutte le stime k-step:

$\hat{A}_t^{GAE(\gamma,\lambda)} = (1-\lambda) \sum_{k=1}^{\infty} \lambda^{k-1} \hat{A}_t^{(k)} = (1-\lambda) \sum_{k=1}^{\infty} \lambda^{k-1} \sum_{l=0}^{k-1} \gamma^l \delta_{t+l} = (1-\lambda) \sum_{l=0}^{\infty} \gamma^l \delta_{t+l} \sum_{k=l+1}^{\infty} \lambda^{k-1} = (1-\lambda) \sum_{l=0}^{\infty} \gamma^l \delta_{t+l} \cdot \lambda^l / (1-\lambda) = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l}$

dove si è scambiato l'ordine di somma e si è usata la serie geometrica $\sum_{k=l+1}^{\infty} \lambda^{k-1} = \lambda^l/(1-\lambda)$. □

La formula ricorsiva $\hat{A}_t^{GAE} = \delta_t + (\gamma\lambda) \hat{A}_{t+1}^{GAE}$ segue immediatamente dalla struttura della somma e permette il calcolo in $O(T)$ scorrendo la traiettoria all'indietro.

Silvano Salvador

A.II.7 — Il lemma di politica migliorata (Kakade-Langford 2002)

Questo lemma è il fondamento teorico di TRPO. Si vuole limitare inferiormente il guadagno $J(\pi) - J(\pi_old)$ per una politica nuova π rispetto a quella vecchia π_old.

Si definisce la distribuzione di visite agli stati (state visitation measure) per un MDP con fattore di sconto γ:

$d^\pi(s) = (1 - \gamma) \sum_{t=0}^{\infty} \gamma^t P(s_t = s \mid \pi)$

con $\sum_s d^\pi(s) = 1$ (misura di probabilità per normalizzazione con il fattore $(1-\gamma)$). L'obiettivo si scrive come:

$J(\pi) = 1/(1-\gamma) \cdot \mathbb{E}_{s \sim d^\pi, a \sim \pi(\cdot \mid s)} [A^{\pi_old}(s, a)] + J(\pi_old)$

Dimostrazione. Si parte dall'identità di telescoping per la funzione di valore:

$V^\pi(s_0) - V^{\pi_old}(s_0) = \mathbb{E}_{\tau \sim \pi} [\sum_{t=0}^{\infty} \gamma^t A^{\pi_old}(s_t, a_t)]$

Questa identità segue espandendo $V^\pi(s_0) = \mathbb{E}_\pi[\sum_t \gamma^t r_t]$ e aggiungendo e sottraendo il valore di π_old:

$V^\pi(s_0) = \mathbb{E}_{\tau \sim \pi} [\sum_t \gamma^t r_t] = \mathbb{E}\{\tau \sim \pi\} [\sum_t \gamma^t (r_t + \gamma V^{\pi_old}(s\{t+1\}) - V^{\pi_old}(s_t)) + \sum_t \gamma^t V^{\pi_old}(s_t) - \sum_t \gamma^{t+1} V^{\pi_old}(s\{t+1\})] = \mathbb{E}\{\tau \sim \pi\} [\sum_t \gamma^t A^{\pi_old}(s_t, a_t)] + V^{\pi_old}(s_0)$

dove la somma telescopica $\sum_t \gamma^t V^{\pi_old}(s_t) - \gamma^{t+1} V^{\pi_old}(s\{t+1\}) = V^{\pi_old}(s_0) - \lim_{T\to\infty} \gamma^T V^{\pi_old}(s_T) = V^{\pi_old}(s_0)$ per $\gamma < 1$ e V^{π_old} limitata. □

Mediando su $s_0 \sim d_0$ e passando alla misura di visite:

Silvano Salvador

$J(\pi) - J(\pi_old) = \mathbb{E}_{s_0 \sim d_0}\,[V^{\pi}(s_0) - V^{\pi_old}(s_0)] = \mathbb{E}_{s_0}\,[\mathbb{E}_{\tau \sim \pi}\,[\sum_t \gamma^t A^{\pi_old}(s_t, a_t)]] = \sum_t \sum_s \gamma^t P(s_t = s \mid \pi, s_0 \sim d_0) \cdot \mathbb{E}\{a \sim \pi(\cdot|s)\}\,[A^{\pi_old}(s, a)] = 1/(1-\gamma) \cdot \mathbb{E}\{s \sim d^{\pi}, a \sim \pi(\cdot|s)\}\,[A^{\pi_old}(s, a)]$

Sostituendo la distribuzione di visite d^{π} con quella di π_old — si introduce un errore controllabile dalla distanza tra le due politiche — e usando l'importance sampling per rimpiazzare l'aspettativa sulle azioni sotto π con una sotto π_old:

$L_{\pi_old}(\pi) := 1/(1-\gamma) \cdot \mathbb{E}_{s \sim d^{\pi_old},\, a \sim \pi_old(\cdot|s)}\,[\pi(a|s)/\pi_old(a|s) \cdot A^{\pi_old}(s, a)]$

è l'obiettivo surrogato di TRPO. Il lemma di Kakade-Langford quantifica l'errore introdotto dalla sostituzione $d^{\pi} \to d^{\pi_old}$. Si definisce la variazione totale massima:

$D^{\max}\{TV\}(\pi, \pi_old) = max_s\, D\{TV\}(\pi(\cdot|s), \pi_old(\cdot|s))$

Allora vale il bound:

$|d^{\pi}(s) - d^{\pi_old}(s)| \le 2\gamma\, D^{\max}_{TV}(\pi, \pi_old) / (1-\gamma)^2$

che, combinato con la limitatezza di A^{π_old}, produce il bound sul guadagno:

$J(\pi) - J(\pi_old) \ge L_{\pi_old}(\pi) - C \cdot (D^{\max}_{TV}(\pi, \pi_old))^2$

con $C = 4\gamma \cdot \max_{s,a} |A^{\pi_old}(s, a)| / (1-\gamma)^2$. Usando $D_{TV} \le \sqrt{(D_{KL}/2)}$ (disuguaglianza di Pinsker):

$J(\pi) - J(\pi_old) \ge L_{\pi_old}(\pi) - C' \cdot D^{\max}_{KL}(\pi_old \,\|\, \pi)$

con $C' = 2C$. Massimizzare questo limite inferiore è l'obiettivo di TRPO.

Silvano Salvador

A.II.8 — TRPO: derivazione dell'aggiornamento e calcolo del gradiente naturale

L'ottimizzazione di TRPO risolve il problema vincolato:

max_{θ} L_{θ_old}(θ) := 𝔼_{s ~ d^{π_{θ_old}}, a ~ π_{θ_old}} [r_t(θ) · Â(s, a)] s.t. 𝔼_s [D_{KL}(π_{θ_old}(· | s) ‖ π_θ(· | s))] ≤ δ

dove r_t(θ) = π_θ(a | s) / π_{θ_old}(a | s). Per risolverlo approssimativamente, si linearizza l'obiettivo e si approssima il vincolo al secondo ordine attorno a θ_old.

Sia Δθ = θ − θ_old. La linearizzazione dell'obiettivo è:

L_{θ_old}(θ_old + Δθ) ≈ L_{θ_old}(θ_old) + g^T Δθ, g = ∇θ *L*{θ_old}(θ) |_{θ = θ_old}

L'espansione al secondo ordine del vincolo KL è:

𝔼_s [D_{KL}(π_{θ_old}(· | s) ‖ π_{θ_old + Δθ}(· | s))] ≈ (1/2) Δθ^T F Δθ

dove F è la matrice di Fisher della politica (che coincide, a secondo ordine, con l'Hessiana della divergenza KL):

F(θ) = 𝔼_{s ~ d^{π_θ}, a ~ π_θ(· | s)} [∇_θ log π_θ(a | s) · (∇_θ log π_θ(a | s))^T]

Questa approssimazione al secondo ordine della KL segue dalla rappresentazione della KL come somma di termini di entropia e cross-entropia; la matrice F ne è la curvatura in θ_old, e vale la relazione:

F = −𝔼_{s ~ d^{π_θ}, a ~ π_θ(· | s)} [∇²_θ log π_θ(a | s)]

(il segno meno riflette la concavità di log π_θ rispetto ai parametri per ogni (s,a) fissato).

Silvano Salvador

Il problema approssimato diventa:

$\max_{\Delta\theta} g^T \Delta\theta$ s.t. $(1/2)\, \Delta\theta^T F \Delta\theta \leq \delta$

Questo è un problema di ottimizzazione quadratica con vincolo quadratico. Applicando le condizioni KKT con moltiplicatore di Lagrange $\lambda \geq 0$:

$$g - \lambda F \Delta\theta = 0 \Longrightarrow \Delta\theta^* = (1/\lambda)\, F^{-1} g$$

Il moltiplicatore λ si determina dal vincolo attivo:

$$(1/2)\, (1/\lambda)^2\, g^T F^{-1} g = \delta \Longrightarrow \lambda = \sqrt{(g^T F^{-1} g / (2\delta))}$$

L'aggiornamento ottimale è:

$$\Delta\theta^* = \sqrt{(2\delta / (g^T F^{-1} g))} \cdot F^{-1} g$$

Il vettore $F^{-1} g$ è il gradiente naturale di Amari: la direzione di massima ascesa dell'obiettivo nella metrica riemanniana di Fisher. L'aggiornamento TRPO è proporzionale al gradiente naturale, con dimensione del passo calibrata in modo che la variazione KL sia esattamente δ. Il fattore di normalizzazione $\sqrt{(2\delta / (g^T F^{-1} g))}$ è l'ampiezza del passo che mantiene la traiettoria sulla superficie $D_{KL} = \delta$ nello spazio dei parametri.

Il calcolo esatto di $F^{-1} g$ richiede l'inversione della matrice $d \times d$ di Fisher, il che è computazionalmente proibitivo per reti neurali con d dell'ordine dei milioni. TRPO lo approssima risolvendo il sistema lineare $F x = g$ con il metodo del gradiente coniugato (CG), che richiede soltanto prodotti matrice-vettore Fv — calcolabili senza formare F esplicitamente mediante la tecnica di Hessian-free optimization:

$$Fv = \nabla_\theta (\nabla_\theta \log \pi_\theta(a|s))^T v \cdot \mathbb{E}[\nabla_\theta \log \pi_\theta(a|s)]$$

Silvano Salvador

In pratica: si calcola il prodotto scalare tra il vettore v e il gradiente del log-politica, se ne prende il gradiente rispetto a θ — due backward pass invece di una — e si ottiene Fv con costo computazionale O(d) anziché $O(d^2)$.

A.II.9 — PPO come approssimazione conservativa di TRPO

L'obiettivo PPO-clip viene qui analizzato nei dettagli, mostrando perché il clipping del rapporto di importanza approssimi il vincolo KL di TRPO.

Si definisca r(θ) = π_θ(a|s) / π_{θ_old}(a|s). Per piccoli aggiornamenti Δθ, vale l'approssimazione:

$D_{KL}(\pi_{\theta_{old}}(\cdot|s) \| \pi_\theta(\cdot|s)) \approx (1/2)\ \mathbb{E}_{a \sim \pi_{\theta_{old}}}[(r(\theta) - 1)^2]$

Questo segue dall'espansione in serie di Taylor di D_{KL} e dal legame tra la distanza in variazione totale e la KL. Il vincolo $D_{KL} \leq \delta$ corrisponde approssimativamente a $\mathbb{E}[(r(\theta) - 1)^2] \leq 2\delta$, ossia $r(\theta) \in [1 - \varepsilon, 1 + \varepsilon]$ con $\varepsilon = \sqrt{(2\delta)}$ per la maggior parte delle azioni.

Il clipping di PPO implementa questo vincolo direttamente sull'obiettivo: invece di penalizzare i valori di r(θ) fuori dall'intervallo [1−ε, 1+ε] attraverso un termine aggiuntivo alla perdita (come nel PPO-KL penalty), li taglia rendendo l'obiettivo insensibile ad aggiornamenti che porterebbero r(θ) fuori dall'intervallo. La funzione clip(r, 1−ε, 1+ε) satura a 1−ε se r < 1−ε e a 1+ε se r > 1+ε.

L'obiettivo $L^{CLIP}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\ \hat{A}_t, \text{clip}(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\ \hat{A}_t)]$ si analizza per casi:

Silvano Salvador

Per $\hat{A}_t > 0$ (azione migliore della media, si vuole aumentare la probabilità): $\min(r_t, \text{clip}(r_t, 1-\varepsilon, 1+\varepsilon)) \cdot \hat{A}_t = \min(r_t, 1+\varepsilon) \cdot \hat{A}_t$

La funzione aumenta linearmente in r_t fino a $r_t = 1+\varepsilon$, poi si satura. Il gradiente $\partial L/\partial r_t = \hat{A}_t \cdot \mathbf{1}[r_t < 1+\varepsilon]$ è zero per $r_t > 1+\varepsilon$: nessun incentivo a spingere r_t oltre la soglia.

Per $\hat{A}_t < 0$ (azione peggiore della media, si vuole diminuire la probabilità): $\min(r_t, \text{clip}(r_t, 1-\varepsilon, 1+\varepsilon)) \cdot \hat{A}_t = \max(r_t, 1-\varepsilon) \cdot \hat{A}_t$

La funzione decresce (diventa più negativa) man mano che r_t diminuisce verso $1-\varepsilon$, poi si satura. Il gradiente $\partial L/\partial r_t = \hat{A}_t \cdot \mathbf{1}[r_t > 1-\varepsilon]$ è zero per $r_t < 1-\varepsilon$: nessun incentivo a spingere r_t sotto la soglia.

In entrambi i casi, il gradiente di L^{CLIP} rispetto a θ si annulla quando l'aggiornamento tenderebbe a portare $r_t(\theta)$ fuori dall'intervallo $[1-\varepsilon, 1+\varepsilon]$, implementando una forma di barriera di gradiente che disincentiva aggiornamenti troppo grandi.

Appendice Tecnica III

Il teorema di impossibilità di Arrow: dimostrazione passo per passo

---

A.III.1 — Il contesto e la struttura del problema

Silvano Salvador

Il teorema di impossibilità di Arrow (1951, 1963) affronta il problema fondamentale dell'aggregazione delle preferenze: date le preferenze di n individui su un insieme di alternative X, è possibile costruire una "volontà collettiva" — un ordinamento sociale — che rispetti criteri di ragionevolezza minimi?

Si lavora nel seguente contesto formale.

Individui. $N = \{1, 2, ..., n\}$ con $n \geq 2$.

Alternative. X con $|X| = m \geq 3$. L'ipotesi $m \geq 3$ è essenziale: per $m = 2$ il maggioritario soddisfa tutte le condizioni di Arrow.

Ordinamento. Un ordinamento (completo e transitivo) su X è una relazione binaria $\succsim \subseteq X \times X$ tale che: per ogni $x, y \in X$, $x \succsim y$ oppure $y \succsim x$ (completezza); per ogni $x, y, z \in X$, se $x \succsim y$ e $y \succsim z$ allora $x \succsim z$ (transitività). Si scrive $x \succ y$ se $x \succsim y$ e $\neg(y \succsim x)$ (preferenza stretta); $x \sim y$ se $x \succsim y$ e $y \succsim x$ (indifferenza).

Lo spazio degli ordinamenti. Sia $\mathscr{P}$ l'insieme di tutti gli ordinamenti su X. Un profilo di preferenze è $R = (R_1, ..., R_n) \in \mathscr{P}^n$: per ogni individuo i, R_i è il suo ordinamento personale su X.

La funzione di benessere sociale (SWF). Una SWF è una funzione $F : \mathscr{P}^n \to \mathscr{P}$ che associa a ogni profilo un ordinamento collettivo F(R).

A.III.2 — Le cinque condizioni di Arrow

Si elencano le condizioni con la loro formalizzazione precisa.

Condizione U (Dominio universale). F è definita su tutto $\mathscr{P}^n$: per ogni profilo $R \in \mathscr{P}^n$, $F(R) \in \mathscr{P}$ è definito e produce un ordinamento completo e transitivo.

Silvano Salvador

Questa condizione è cruciale: esclude le SWF definite solo su sottodomini ristretti di profili (come il dominio single-peaked di Black che permette il voto a maggioranza per certe strutture di preferenze).

Condizione P (Pareto debole). Per ogni $R \in \mathscr{P}^h$ e ogni $x, y \in X$:

$[x >_i y \text{ per ogni } i \in N] \Rightarrow [x >_{F(R)} y]$

Se tutti gli individui preferiscono strettamente x a y, la collettività preferisce x a y.

Condizione IIA (Indipendenza dalle alternative irrilevanti). Per ogni $R, R' \in \mathscr{P}^h$ e ogni $x, y \in X$:

$[\forall i: (x\ R_i\ y \Leftrightarrow x\ R'_i\ y) \wedge (y\ R_i\ x \Leftrightarrow y\ R'_i\ x)] \Rightarrow [x\ F(R)\ y \Leftrightarrow x\ F(R')\ y]$

Il confronto collettivo tra x e y dipende esclusivamente dai confronti individuali tra x e y.

Condizione ND (Non-dittatorialità). Non esiste $d \in N$ tale che per ogni $R \in \mathscr{P}^h$ e ogni $x, y \in X$:

$x >_d y \Rightarrow x >_{F(R)} y$

Nessun individuo impone le proprie preferenze al collettivo su ogni coppia.

Condizione T (Razionalità collettiva). $F(R) \in \mathscr{P}$ per ogni R — l'ordinamento collettivo è completo e transitivo. Questa è già inclusa nella definizione di SWF ma vale la pena renderla esplicita come requisito separato.

A.III.3 — Definizioni chiave: decisività e quasi-decisività

Silvano Salvador

Per la dimostrazione, le nozioni di decisività sono centrali.

Definizione 1 (Quasi-decisività). La coalizione $C \subseteq N$ è quasi-decisiva sulla coppia (x, y) — indicato $C \in QD(x, y)$ — se esiste almeno un profilo $\vec{R} \in \mathscr{P}^n$ tale che: (i) $x \succ_i y$ per ogni $i \in C$ (ii) $y \succ_i x$ *per ogni* $i \notin C$ *(iii)* $x \succ \{F(\vec{R})\}$ y

Ovvero: C è quasi-decisiva se, in almeno un profilo dove C vuole x e tutti gli altri vogliono y, vince x.

Definizione 2 (Decisività). La coalizione $C \subseteq N$ è decisiva sulla coppia (x, y) — indicato $C \in D(x, y)$ — se per ogni $R \in \mathscr{P}^n$:

$[x \succ_i y$ *per ogni* $i \in C] \Longrightarrow [x \succ \{F(R)\}\ y]$

Ovvero: C è decisiva se, in qualsiasi profilo dove tutti in C preferiscono x a y, il collettivo preferisce x a y — indipendentemente da come si schierano gli altri.

Osservazione. Ogni coalizione decisiva è quasi-decisiva (prendendo il profilo richiesto come uno qualunque tra quelli in cui C preferisce x a y); non è vero il contrario in generale. Il Lemma 1 dimostrerà che, sotto U, P, IIA e T, quasi-decisività implica decisività su tutte le coppie.

A.III.4 — Lemma 1: espansione del campo di decisività

Lemma 1 (Field Expansion Lemma). Se una coalizione C è quasi-decisiva sulla coppia (x, y) per qualche $x \neq y \in X$, allora C è decisiva su ogni coppia (a, b) con $a, b \in X$, $a \neq b$.

La dimostrazione si articola in quattro passi che coprono tutti i casi relativi al posizionamento di (a, b) rispetto a (x, y).

Silvano Salvador

Passo 1: C è decisiva su (x, y).

Sia R^- il profilo in cui $C \in QD(x, y)$. Si vuole mostrare che, per qualsiasi profilo R in cui $x >_i y$ *per ogni* $i \in C$, *vale* $x >_{F(R)} y$ — indipendentemente dalle preferenze di $N \setminus C$ su (x, y).

Sia R qualsiasi profilo con $x >_i y$ per ogni $i \in C$. Per U, esiste un profilo R* che coincide con R sui confronti (x, y) di tutti i giocatori, ma dove le preferenze sui confronti non-coinvolgenti x o y sono libere. Per IIA, il confronto F(R) vs F(R^-) su (x, y) dipende solo dai confronti individuali su (x, y). In R^-: C preferisce x a y, N\C preferisce y a x. In R: C preferisce x a y, N\C ha preferenze qualsiasi su (x, y). Per IIA, il confronto F(R) su (x, y) dipende solo dai confronti individuali su (x, y). Ma stiamo usando la quasi-decisività su un profilo specifico...

Il passaggio preciso richiede un profilo ausiliario. Sia R qualsiasi profilo con $x >_i y \ \forall i \in C$. Costruiamo R* tale che: su (x,y), coincide con R; per $i \in C$, l'ordinamento completo mette x al primo posto ($x >_{i^*}$ tutto); per $j \notin C$, l'ordinamento completo mette y al primo posto. Per Pareto applicato a coppie non-xy: l'ordinamento F(R*) è consistente. Per il profilo R^- della quasi-decisività, i confronti su (x,y) di C e N\C sono gli stessi di R* ristretti a (x,y)? Non necessariamente, perché R^- ha N\C con $y > x$, mentre R* può avere N\C con qualsiasi comparazione.

Il modo corretto è il seguente argomento diretto: la quasi-decisività dice che esiste R^- con $x >_i y \ \forall i \in C$, $y >_i x \ \forall i \notin C$, *e* $x >_{F(R^-)} y$. *Per IIA, F(R) su (x,y) dipende solo da chi preferisce* x *a* y *e chi preferisce* y *a* x *in* R. *Definiamo la funzione di aggregazione binaria* $f_{xy}: \{0,1\}^n \to \{0,1\}$ dove $f_{xy}(b) = 1$ iff $x >_{F(R)} y$, *con* $b_i = 1$ *iff* $x >_i y$. *Per IIA, questa funzione è ben definita (non dipende dalle preferenze su coppie diverse da (x,y)). Il profilo* R^- *ha* $b = 1_C$ *(il vettore indicatore di C) e* $f_{xy}(1_C) = 1$. *Per P,* $f_{xy}(1_N) = 1$. Per qualsiasi b con $b_i = 1 \ \forall i \in C$ (indipendentemente da b per $j \notin C$),

Silvano Salvador

vogliamo mostrare f_{xy}(b) = 1. Per il lemma di monotonia (che segue da P e IIA), se si passa da un profilo dove x $>_i$ y *per* $i \in C$ *e* y $>_i$ x *per* $i \notin C$ *a uno dove* $i \in C$ *mantengono* $x > y$ *e qualcuno in* $N \setminus C$ *cambia da* $y > x$ *a* $x > y$, *il risultato non può cambiare da* $x >$ {F} *y a* $y >$ {F} x. Questo segue perché: prendere il profilo $\bar{R}$ e modificare solo i giocatori $j \notin C$ da $y > x$ a $x > y$ è coerente con Pareto (nessuno ha cambiato in direzione contraria). Formalmente si richiede un argomento di monotonia (estensione lemma) che qui si assume noto. Quindi $C \in D(x, y)$.

Passo 2: Da $C \in D(x, y)$ a $C \in D(x, z)$ per $z \neq y$, $z \neq x$.

Sia $z \in X \setminus \{x, y\}$. Si costruisce il profilo R_1:

- per ogni $i \in C$: $x >_{1_i} y$, $x >_{1_i} z$, $y >_{1_i} z$ (ovvero $x >_1 y >_1 z$)
- per ogni $j \notin C$: $y >_{1_j} z$, $z >_{1_j} x$ (e $y >_{1_j} x$, ottenibile per transitività da $y > z > x$)

Poiché $x >_{1_i} y$ per ogni $i \in C$ e $C \in D(x, y)$, si ha $x >$ {*F*(R_1)} *y. Poiché* $y >_{1_j} z$ *per ogni* $j \in N$ *(sia C che* $N \setminus C$ *hanno* $y > z$ *in* R_1*), per Pareto P:* $y >$ {F(R_1)} z. Per transitività di F(R_1): $x >_{_}$ {F(R_1)} z.

Ora le preferenze su (x, z) in R_1 sono: $x >_{_i} z$ per ogni $i \in C$ (dalla transitività $x > y > z$); $z >_{_j} x$ per ogni $j \notin C$. Per IIA, F(R_1) su (x, z) dipende solo dai confronti individuali su (x, z). Per qualunque profilo R con $x >_i z$ $\forall i \in C$ *(indipendentemente dalle preferenze di* $N \setminus C$ *su (x,z)), il confronto su (x,z) in R ha gli stessi segnali di* R_1 *per* $i \in C$. *Lo stesso argomento di monotonia di prima implica* $x >$ {F(R)} z. Quindi $C \in D(x, z)$.

Passo 3: Da $C \in D(x, y)$ a $C \in D(z, y)$ per $z \neq x$, $z \neq y$.

Sia $z \in X \setminus \{x, y\}$. Si costruisce il profilo R_2:

- per ogni $i \in C$: $z >_{2_i} x >_{2_i} y$ (ovvero $z > x > y$)

Silvano Salvador

- per ogni $j \notin C$: $y >_{2_j} z$ e $x >_{2_j} z$ (con ordine $y > x > z$ oppure $y > z$ e $x > z$ separatamente)

Più precisamente per $j \notin C$: $y >_{2_j} x >_{2_j} z$ (per avere transitività).

Poiché $x >_{2_i} y$ per ogni $i \in C$ e $C \in D(x, y)$: $x >_{_\{F(R_2)\}} y$. Poiché $z >_{2_i} x$ per ogni $i \in C$ e per ogni $j \notin C$: $z >_{2_j} x$? No: $j \notin C$ ha $y > x > z$, dunque $x >_{_j} z$, non $z >_{_j} x$.

Si correggono gli ordinamenti: per $i \in C$ la sequenza è $z > x > y$, quindi $z >_{_i} x$ e $x >_{_i} y$ e $z >_{_i} y$. Per $j \notin C$ sia $y >_{_j} z >_{_j} x$: quindi $z >_{_j} x$ per ogni $j \notin C$.

Ora per il confronto (z, x): $z >_{_i} x \ \forall i \in C$ e $x >_{_j}$... no, abbiamo $z >_j x$ *per* $j \notin C$ *con l'ordinamento* $y > z > x$. *Dunque tutti in N hanno* $z > x$, *e per Pareto* $z >_{\{F(R_2)\}} x$.

Poiché $x >_i y$ *per* $i \in C$ *e* $C \in D(x,y)$: $x >_{\{F(R_2)\}} y$. Per transitività di $F(R_2)$: $z >_{_\{F(R_2)\}} y$.

I confronti su (z, y) in R_2: $z >_{_i} y \ \forall i \in C$ (poiché $z > x > y$); $y >_{_j} z \ \forall j \notin C$. Per IIA e argomento di monotonia: $C \in D(z, y)$.

Passo 4: Da $C \in D(x, y)$ a $C \in D(a, b)$ per qualsiasi (a, b) con $a \neq b$.

Se $a \notin \{x, y\}$ e $b \notin \{x, y\}$: dal Passo 2, $C \in D(x, b)$; dal Passo 3 applicato a (x, b) con $z = a$, $C \in D(a, b)$. Se $a = x$ e $b \notin \{x, y\}$: Passo 2 direttamente. Se $a \notin \{x, y\}$ e $b = y$: Passo 3 direttamente. Se $a = y$ e $b = x$: $C \in D(x, b)$ con b = qualsiasi terza alternativa z (Passo 2); poi $C \in D(a, z)$ con $a = y$ (Passo 3); poi $C \in D(y, x)$ per combinazione. I rimanenti casi si coprono per simmetria degli argomenti.

Quindi C è decisiva su ogni coppia di alternative. □

Silvano Salvador

A.III.5 — Lemma 2: contrazione della coalizione decisiva

Lemma 2 (Group Contraction Lemma). Se una coalizione C con $|C| \geq 2$ è decisiva su qualche coppia — equivalentemente, su ogni coppia per il Lemma 1 — allora esiste una sottocoalizione propria $C' \subsetneq C$, $C' \neq \varnothing$, che è quasi-decisiva su qualche coppia.

Dimostrazione. Siano $i, j \in C$ due individui distinti. Definiamo $C_1 = \{i\}$ e $C_2 = C \setminus \{i\}$. Si prendano tre alternative $x, y, z \in X$ distinte (possibile per $|X| \geq 3$).

Costruiamo il profilo R_3:

- individuo i (in C_1): $x >_{3_i} y >_{3_i} z$
- individui $k \in C_2$: $y >_{3_k} z >_{3_k} x$
- individui $j \notin C$: $z >_{3_j} x >_{3_j} y$

(Si verifica che questi ordinamenti sono completi e transitivi per costruzione.)

Si considera il confronto $F(R_3)$ su (y, z). Poiché $y >_{3_k} z$ per ogni $k \in C_2$ (e $y >_{3_i} z$ per l'individuo i), tutti i membri di C hanno $y > z$ in R_3. Per decisività di C su (y, z): $y >_\{F(R_3)\} z$.

Ora si considerano i due casi possibili per $F(R_3)$ su (x, y):

Caso A: $x >_\{F(R_3)\} y$.

Allora per transitività: $x >\{F(R_3)\} y >\{F(R_3)\} z$, cioè $x >_\{F(R_3)\} z$.

Nel profilo R_3, i confronti su (x, z) sono: $x >_{3_i} z$ (poiché $x > y > z$ per i); $z >_{3_j} x$ per ogni $j \notin C$ (dato che $z > x$ per i non-C); $z >_{3_k} x$ per ogni $k \in C_2$ (dato che $y > z > x$).

Silvano Salvador

Quindi il solo individuo che ha $x \succ z$ è $i = C_1$, mentre tutti gli altri hanno $z \succ x$. Eppure il collettivo ha $x \succ z$. Questo è esattamente la definizione di $C_1 = \{i\}$ quasi-decisivo su (x, z): nel profilo R_3, solo i vuole x sopra z e vince. Dunque $\{i\} \in QD(x, z) \subsetneq C$.

Caso B: $y \succ_{F(R_3)} x$ *o* $y \sim_{F(R_3)} x$.

Nel profilo R_3, i confronti su (x, y) sono: $x \succ_{3,i} y$ per l'individuo i; $y \succ_{3,k} x$ per ogni $k \in C_2$; $x \succ_{3,j} y$ per ogni $j \notin C$ (dato che $x \succ y$ per $j \notin C$). In questo profilo, hanno $x \succ y$: l'individuo i e tutti i $j \notin C$. Hanno $y \succ x$: i $k \in C_2$.

Poiché il collettivo ha $y \succeq_{F(R_3)} x$ (Caso B), e la coalizione C_2 è l'unico gruppo che preferisce y a x mentre tutti gli altri (i e $N \setminus C$) preferiscono x a y, la coalizione C_2 deve essere quasi-decisiva su (y, x): $C_2 \in QD(y, x) \subsetneq C$.

In entrambi i casi, una sottocolalizione propria di C è quasi-decisiva su qualche coppia. □

Nota: il Caso A e il Caso B non si escludono a vicenda quando $F(R_3)$ produce $x \sim_{F(R_3)} y$ *(indifferenza). In quel caso, per transitività con* $y \succ_{F(R_3)} z$, si ha $x \succeq_{F(R_3)} z$, e il Caso A si applica ugualmente.

A.III.6 — Dimostrazione del teorema di Arrow

Teorema (Arrow, 1951). Non esiste una SWF F che soddisfi simultaneamente U, P, IIA e ND con $|X| \geq 3$, $n \geq 2$, e tale che F(R) sia un ordinamento completo e transitivo per ogni R.

Dimostrazione. Si supponga per assurdo che esista una tale F.

Passo 1: La coalizione N di tutti gli individui è decisiva su ogni coppia.

Silvano Salvador

Per la condizione P, se $x \succ_i y$ *per ogni* $i \in N$, *allora* $x \succ \{F(R)\}$ y. Quindi $N \in D(x, y)$ per ogni (x, y). In particolare N è una coalizione decisiva.

Passo 2: Costruzione della catena di coalizioni decrescenti.

Per il Lemma 2 applicato a N (che soddisfa $|N| \geq 2$): esiste $C_1 \subsetneq N$ con $|C_1| < |N|$, $C_1 \in QD$ su qualche coppia. Per il Lemma 1: $C_1 \in D$ su ogni coppia.

Si applica iterativamente il Lemma 2: C_1 è decisiva con $|C_1| \geq 2 \Longrightarrow$ esiste $C_2 \subsetneq C_1$, $C_2 \in QD \Longrightarrow C_2 \in D$ per Lemma 1 C_2 è decisiva con $|C_2| \geq 2 \Longrightarrow$ esiste $C_3 \subsetneq C_2$, $C_3 \in QD \Longrightarrow C_3 \in D$ per Lemma 1 ...

Poiché $|C_{+1}| < |C|$ e le cardinalità sono interi positivi, la catena termina necessariamente: esiste un k* tale che $|C_\{k^*\}| = 1$.

Passo 3: Il singoletto è una dittatura.

Sia $C_\{k^*\} = \{d\}$ con un unico individuo d. Per il Lemma 1, {d} è decisiva su ogni coppia (x, y): se $x \succ_d y$, *allora* $x \succ \{F(R)\}$ y per qualsiasi R.

Questa è esattamente la definizione di dittatura di d — la condizione esclusa da ND. Contraddizione.

Quindi non può esistere una SWF che soddisfi tutte le condizioni. □

A.III.7 — Analisi delle condizioni: cosa succede rilassando ciascuna

La comprensione profonda del teorema si acquisisce analizzando cosa accade quando si rilassa ciascuna delle cinque condizioni.

Rilassando U (dominio universale). Si consente che F sia definita solo su un sottodominio di profili. Il voto a maggioranza (majority rule) soddisfa P, IIA e ND, ma viola U: produce ordinamenti transitivi solo quando le

Silvano Salvador

preferenze degli individui sono single-peaked (ordinabili su una linea reale senza "selle"). Questo è il teorema del mediano di Black (1948).

Rilassando P (Pareto). Si ottengono funzioni come la regola costante (F(R) = R_fixed per un ordinamento fisso) che soddisfano IIA, ND, U e T, ma violano P perché non rispecchiano unanimità.

Rilassando IIA. Si apre la porta alle regole di Borda (Borda count), alla regola di Condorcet, ai metodi di votazione preferenziale — tutti non-dittatoriali, Pareto, con dominio universale — ma nessuno di essi produce un ordinamento collettivo transitivo in tutti i casi (paradosso di Condorcet).

Rilassando ND. Si ottiene la dittatura: F(R) = R_d per un individuo fisso d soddisfa U, P, IIA e T, ma è dittatoriale.

Rilassando T (razionalità collettiva, ovvero transitività dell'ordinamento collettivo). Se si permette che F(R) sia soltanto una relazione di preferenza aciclidica (non necessariamente transitiva), esistono SWF non-dittatoriali che soddisfano U, P e IIA — per esempio la maggioranza semplice. Il prezzo è la possibilità di cicli (A > B > C > A) che rendono l'ordinamento collettivo incoerente.

Rilassando la cardinalità delle alternative (|X| = 2). Con solo due alternative, il voto a maggioranza soddisfa tutte e cinque le condizioni — è il teorema di May (1952) per due alternative.

A.III.8 — Connessione con il teorema di Gibbard-Satterthwaite

Il teorema di Gibbard-Satterthwaite (1973, 1975) è il duale del teorema di Arrow nel contesto dei meccanismi di implementazione: afferma che

Silvano Salvador

non esiste un meccanismo di scelta sociale (una funzione che mappa i profili di preferenze dichiarate in un'alternativa) che sia simultaneamente:

onto (surjettivo: ogni alternativa può essere scelta), strategy-proof (per ogni individuo, dire la verità è debolmente dominante), non-dittatoriale.

Formalmente, sia $g : \mathscr{P}^n \to X$ una funzione di scelta sociale. La strategy-proofness richiede:

$\forall i \in N, \forall R \in \mathscr{P}^n, \forall R'_i \in \mathscr{P}: g(R_i, R_{-i}) \; R_i \; g(R'_i, R_{-i})$

ovvero dichiarare la vera preferenza R_i porta a un risultato almeno buono quanto qualsiasi dichiarazione alternativa R'_i. Il teorema afferma che qualsiasi g onto e strategy-proof deve essere una dittatura — esiste d tale che g(R) = top(R_d) per ogni R (il risultato è sempre la scelta migliore del dittatore).

La connessione con Arrow segue dalla struttura della dimostrazione: entrambi i teoremi usano argomenti di decisività e contrazione delle coalizioni per ridurre il potere decisionale a un singolo individuo. Gibbard-Satterthwaite si ricava direttamente da Arrow imponendo che la scelta collettiva sia deterministica (un'alternativa, non un ordinamento) e considerando le implicazioni dell'incentive-compatibility.

La rilevanza per l'allineamento è che ogni meccanismo per estrarre le vere preferenze degli utenti e tradurle in comportamento di un sistema di AI è soggetto alla stessa impossibilità: o è manipolabile (gli utenti hanno incentivo a dichiarare preferenze distorte), o è dittatoriale (le preferenze di un solo individuo o gruppo determinano il comportamento), o non copre tutte le alternative possibili. Non esiste una via di uscita formale da questo trilemma — soltanto compromessi pratici consapevoli.

Appendice Tecnica IV

Silvano Salvador

Controllo ottimo continuo, equazione di Hamilton-Jacobi-Bellman e reinforcement learning con regolarizzazione entropica

A.IV.1 — Il passaggio al continuo: motivazioni e struttura

La teoria degli MDP sviluppata nelle parti precedenti opera in tempo discreto e su spazi di stato e azione che, per ragioni di trattabilità analitica, si sono spesso assunti finiti o al più numerabili. Il mondo fisico è però continuo in quasi ogni sua manifestazione rilevante per il controllo: la posizione e la velocità di un robot, le variabili di stato di un processo chimico, le allocazioni continue di un portafoglio finanziario. Estendere la teoria alle versioni a tempo continuo e a spazio continuo non è soltanto un esercizio di generalizzazione formale: rivela strutture matematiche più profonde — la connessione con il calcolo delle variazioni, con la meccanica hamiltoniana e con le equazioni differenziali alle derivate parziali — che nell'analisi discreta rimangono oscurate.

Il collegamento tra la programmazione dinamica discreta e la teoria classica del controllo ottimo fu stabilito in modo definitivo da Bellman stesso negli anni cinquanta, attraverso un processo di limite che porta dall'equazione di ricorrenza discreta all'equazione differenziale alle derivate parziali nota come Hamilton-Jacobi-Bellman. Questo limite illumina retrospettivamente la struttura dell'operatore di Bellman: la contrazione T^* nello spazio discreto è l'analogo dell'operatore differenziale del primo ordine nell'equazione HJB nello spazio continuo,

Silvano Salvador

e la funzione di valore ottimale V* è in entrambi i casi il punto fisso di questo operatore.

A.IV.2 — Il problema del controllo ottimo deterministico

Si considera il sistema dinamico deterministico descritto dall'equazione differenziale ordinaria:

$$\dot{x}(t) = f(x(t), u(t)),\ x(t) \in \mathbb{R}^n,\ u(t) \in U \subseteq \mathbb{R}^m,\ t \in [0, T]$$

dove x(t) è il vettore di stato, u(t) è il controllo (l'azione dell'agente), $f : \mathbb{R}^n \times U \to \mathbb{R}^n$ è il campo vettoriale che descrive le dinamiche, e U è l'insieme ammissibile dei controlli. Il funzionale obiettivo è:

$$J(x_0, u(\cdot)) = \int_0^T L(x(t), u(t))\, dt + \Phi(x(T))$$

dove $L : \mathbb{R}^n \times U \to \mathbb{R}$ è la funzione di costo istantaneo (l'analogo di $-R$ nell'MDP, con segno invertito poiché si minimizza), $\Phi : \mathbb{R}^n \to \mathbb{R}$ è il costo terminale e $x_0 = x(0)$ è lo stato iniziale. Il problema del controllo ottimo è trovare il controllo $u^*(\cdot) : [0, T] \to U$ che minimizza J.

La funzione valore ottimale è:

$$V(x, t) = \min_{u(\cdot) \in \mathcal{U}} J(x, u(\cdot))$$

dove $\mathcal{U}$ è lo spazio dei controlli ammissibili (funzioni misurabili $u : [t, T] \to U$) e la minimizzazione è soggetta alle dinamiche $\dot{x} = f(x, u)$ con $x(t) = x$. Per il principio di ottimalità di Bellman, V soddisfa la proprietà di sotto-struttura ottimale:

$$V(x, t) = \min_{u(\cdot) \text{ su } [t,\ t+\Delta t]} \left\{ \int_t^{t+\Delta t} L(x(s), u(s))\, ds + V(x(t+\Delta t), t+\Delta t) \right\}$$

Silvano Salvador

A.IV.3 — Derivazione dell'equazione di Hamilton-Jacobi-Bellman

Si sviluppa V(x(t+Δt), t+Δt) in serie di Taylor al primo ordine in Δt:

$$V(x(t+\Delta t), t+\Delta t) = V(x(t), t) + \partial_t V(x(t), t) \cdot \Delta t + \nabla_x V(x(t), t)^T \cdot \Delta x + o(\Delta t)$$

dove $\Delta x = x(t+\Delta t) - x(t) = f(x(t), u(t)) \cdot \Delta t + o(\Delta t)$ per le dinamiche continue. Sostituendo e approssimando $\int_t^{t+\Delta t} L(x(s), u(s))\, ds \approx L(x(t), u(t)) \cdot \Delta t$:

$$V(x, t) \approx \min_{u \in U} \{L(x, u) \cdot \Delta t + V(x, t) + \partial_t V \cdot \Delta t + \nabla_x V^T f(x, u) \cdot \Delta t + o(\Delta t)\}$$

Sottraendo V(x, t) da entrambi i lati, dividendo per Δt e portando $\Delta t \to 0$:

$$0 = \partial t\, V(x, t) + \min\{u \in U\} \{L(x, u) + \nabla_x V(x, t)^T f(x, u)\}$$

con la condizione al contorno $V(x, T) = \Phi(x)$. Questa è l'equazione di Hamilton-Jacobi-Bellman (HJB) in forma di minimizzazione. Introducendo l'Hamiltoniano del controllo:

$$H(x, p, u) = L(x, u) + p^T f(x, u)$$

dove $p = \nabla_x V$ è il vettore costato (costate vector) o variabile aggiunte, l'equazione HJB diventa:

$$\partial t\, V(x, t) + \min\{u \in U\}\, H(x, \nabla_x V, u) = 0,\ V(x, T) = \Phi(x)$$

e l'Hamiltoniano ottimale è:

$$H^*(x, p) = \min_{u \in U} H(x, p, u) = \min_{u \in U} \{L(x, u) + p^T f(x, u)\}$$

Il controllo ottimale è:

Silvano Salvador

$u^*(x, t) = \text{argmin}_{u \in U} H(x, \nabla_x V(x, t), u)$

e sotto opportune condizioni di regolarità (convessità di H in u, lipschitzianità di f e L) questo argmin è unico e produce un feedback control — un controllo che dipende dallo stato corrente e dal tempo, non dall'intera storia.

A.IV.4 — Il principio del massimo di Pontryagin

Il principio del massimo di Pontryagin (PMP, Pontryagin-Boltyansky-Gamkrelidze-Mishchenko, 1962) fornisce le condizioni necessarie di ottimalità per il problema del controllo ottimo senza richiedere la differenziabilità di V — una proprietà importante perché, anche per problemi regolari, V può mancare di derivate classiche in insiemi di misura nulla. Il PMP opera nello spazio delle traiettorie anziché nello spazio delle funzioni valore, ed è fondato sul calcolo delle variazioni e sulla teoria dei moltiplicatori di Lagrange per sistemi dinamici.

Si introduce il vettore costato $\lambda(t) \in \mathbb{R}^n$ — variabile aggiunta al sistema — e l'Hamiltoniano di Pontryagin:

$\mathcal{H}(x, \lambda, u) = -L(x, u) + \lambda^T f(x, u)$

(con segno opposto rispetto all'Hamiltoniano HJB per la convenzione di massimizzazione). Il PMP afferma che, se $u^*(\cdot)$ è un controllo ottimale con traiettoria corrispondente $x^*(\cdot)$, allora esiste un vettore costato $\lambda^*(\cdot) : [0, T] \to \mathbb{R}^n$ tale che:

Equazione del sistema (stato): $\dot{x}^*(t) = \partial_\lambda \mathcal{H} = f(x^*(t), u^*(t))$ Equazione aggiunta (costato): $\lambda^*(t) = -\partial_x \mathcal{H} = \nabla_x L(x^*(t), u^*(t)) - (\partial_x f)^T \lambda^*(t)$

Silvano Salvador

Condizione di massimalità: u(t) = argmax*{u ∈ U} $\mathcal{H}$(x*(t), λ*(t), u) q.o. su [0, T] Condizioni al contorno: x*(0) = x_0, λ*(T) = ∇_x Φ(x*(T))

Le equazioni del sistema e dell'aggiunta formano un sistema hamiltoniano nel senso della meccanica classica: (ẋ, λ) = (∂_λ $\mathcal{H}$, −∂_x $\mathcal{H}$). La struttura è simplettica: lo spazio (x, λ) ∈ $\mathbb{R}^{2n}$ è lo spazio delle fasi, equipaggiato con la forma simplettica ω = dλ ∧ dx, e il flusso hamiltoniano preserva questa forma (teorema di Liouville per i sistemi hamiltoniani).

La connessione tra il PMP e l'HJB è la seguente. Se V è differenziabile, allora λ*(t) = ∇_x V(x*(t), t) lungo la traiettoria ottimale: il vettore costato è il gradiente spaziale della funzione valore. Le equazioni di costato del PMP corrispondono alle equazioni delle caratteristiche dell'HJB — le curve lungo le quali l'equazione differenziale parziale si riduce a un sistema di ODE. Il controllo ottimale u* è la stessa funzione sia nel PMP (massimalità di $\mathcal{H}$) sia nell'HJB (minimalità di H con p = ∇_x V).

A.IV.5 — L'HJB stocastico: il lemma di Itô e la diffusione

Nel reinforcement learning, le dinamiche dell'ambiente sono tipicamente stocastiche. Il framework naturale per il controllo ottimo stocastico è quello delle equazioni differenziali stocastiche (SDE) di Itô. Si consideri il processo stocastico controllato:

dx(t) = f(x(t), u(t)) dt + σ(x(t), u(t)) dW(t)

dove W(t) ∈ $\mathbb{R}^d$ è un moto browniano standard (processo di Wiener), σ : $\mathbb{R}^n$ × U → $\mathbb{R}$^{n×d} è la matrice di diffusione, e il prodotto σ dW è il termine di rumore stocastico. Il funzionale obiettivo è ora il costo atteso:

Silvano Salvador

$J(x, t, u(\cdot)) = \mathbb{E}\,[\int_t^T L(x(s), u(s))\,ds + \Phi(x(T)) \mid x(t) = x]$

e la funzione valore è $V(x, t) = \min_{u(\cdot)} J(x, t, u(\cdot))$.

Per derivare l'HJB stocastico, si applica il lemma di Itô alla funzione $V(x(t), t)$ che, per ipotesi, è due volte differenziabile in x e una volta in t:

$dV = \partial_t V\,dt + \nabla_x V^T\,dx + (1/2)\,\mathrm{tr}(\sigma\,\sigma^T\,\mathrm{Hess}_x V)\,dt = [\partial_t V + \nabla_x V^T f + (1/2)\,\mathrm{tr}(\sigma\,\sigma^T\,\mathrm{Hess}_x V)]\,dt + \nabla_x V^T\,\sigma\,dW$

Il termine $\mathrm{tr}(\sigma\,\sigma^T\,\mathrm{Hess}_x V) = \sum\{i,j\}\,(\sigma\sigma^T)\{ij\}\,\partial^2 V/\partial x_i \partial x$ è il termine di diffusione del secondo ordine, assente nel caso deterministico. Applicando il principio di ottimalità di Bellman e prendendo il limite $\Delta t \to 0$ (con le aspettative che eliminano il termine di martingala $\nabla_x V^T\,\sigma\,dW$):

$\partial t\,V + \min\{u \in U\}\,\{L(x, u) + \nabla_x V^T f(x, u) + (1/2)\,\mathrm{tr}(\sigma(x,u)\,\sigma(x,u)^T\,\mathrm{Hess}_x V)\} = 0$

con $V(x, T) = \Phi(x)$. Questa è l'equazione di Hamilton-Jacobi-Bellman stocastica. Il termine $(1/2)\,\mathrm{tr}(\sigma\sigma^T\,\mathrm{Hess}_x V)$ è l'operatore di diffusione del secondo ordine, analogo al laplaciano per σ isotropa e costante.

A.IV.6 — Dall'HJB all'MDP: il limite del tempo discreto

La connessione tra l'HJB e l'equazione di Bellman discreta si ottiene attraverso la discretizzazione temporale. Sia Δt un passo temporale piccolo. Approssimando le dinamiche continue con l'integrazione di Eulero stocastica:

$x_{k+1} = x_k + f(x_k, a_k)\,\Delta t + \sigma(x_k, a_k)\,\sqrt{\Delta t} \cdot \varepsilon_k,\ \varepsilon_k \sim N(0, I)$

Silvano Salvador

e il costo istantaneo con $R(x_k, a_k) = -L(x_k, a_k)\,\Delta t$, si ottiene un MDP con fattore di sconto $\gamma = 1 - \lambda\,\Delta t$ (per un tasso di sconto $\lambda > 0$ continuo) e ricompensa R. L'equazione di Bellman discreta per V^* converge all'HJB quando $\Delta t \to 0$, con l'Hessiana $\mathrm{Hess}_x V$ che emerge come correzione del secondo ordine.

Questa derivazione chiarisce anche il significato del fattore di sconto γ: esso corrisponde al tasso di sconto esponenziale continuo $e^{-\lambda t}$ che rende finanziariamente o biologicamente ragionevole preferire ricompense presenti a ricompense future.

A.IV.7 — Il reinforcement learning con regolarizzazione entropica

Il framework del controllo ottimo stocastico si connette al reinforcement learning moderno attraverso il concetto di regolarizzazione entropica della politica. Si considera l'obiettivo modificato:

$$J_{ent}(\pi) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) + \alpha\, H(\pi(\cdot \mid s_t))\right]\right.$$

dove $H(\pi(\cdot \mid s)) = -\sum_a \pi(a \mid s) \log \pi(a \mid s)$ è l'entropia della politica allo stato s e $\alpha > 0$ è il coefficiente di temperatura che controlla il peso relativo di ricompensa e regolarizzazione entropica. Il termine di entropia incoraggia politiche distribuite — esplora le azioni invece di colassare su deterministic policies — e ha un'interpretazione bayesiana precisa: esso implementa il principio di massima entropia di Jaynes, mantenendo la politica la più incerta possibile compatibilmente con la massimizzazione della ricompensa.

L'obiettivo J_{ent} viene detto a volte free energy objective per la sua analogia con l'energia libera di Helmholtz della meccanica statistica: $F =$

Silvano Salvador

$-kT \log Z$ dove Z è la funzione di partizione. In questo formalismo, la politica ottimale è la distribuzione di Gibbs sull'azione:

$\pi^*(a \mid s) \propto \exp(Q^*(s, a) / \alpha)$

Il soft Q-value $Q^*(s, a)$ (distinto dal Q-value standard) soddisfa la soft Bellman equation:

$Q^*(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot \mid s,a)} [V^*(s')]$

dove la soft value function è:

$V^*(s) = \alpha \log \sum_a \exp(Q^*(s, a) / \alpha) = \mathbb{E}_{a \sim \pi^*(\cdot \mid s)} [Q^*(s, a)] + \alpha H(\pi^*(\cdot \mid s))$

Si noti che $V^*(s) = \alpha \log \sum_a \exp(Q^*(s,a)/\alpha)$ è il log-sum-exp smussato del Q-value — la versione morbida (soft) del max che si riduce al max per $\alpha \to 0$ e all'aspettativa uniforme per $\alpha \to \infty$.

La derivazione della politica ottimale $\pi^*(a \mid s) \propto \exp(Q^*(s,a)/\alpha)$ segue dalla condizione KKT del problema di massimizzazione:

$\max_{\pi(\cdot \mid s)} \mathbb{E}_{a \sim \pi} [Q(s, a)] + \alpha H(\pi(\cdot \mid s))$ s.t. $\sum_a \pi(a \mid s) = 1$

Il lagrangiano è:

$\mathcal{L}(\pi, \lambda) = \sum_a \pi(a \mid s) Q(s,a) - \alpha \sum_a \pi(a \mid s) \log \pi(a \mid s) - \lambda(\sum_a \pi(a \mid s) - 1)$

Ponendo $\partial \mathcal{L} / \partial \pi(a \mid s) = 0$:

$Q(s,a) - \alpha(\log \pi(a \mid s) + 1) - \lambda = 0$ $\log \pi(a \mid s) = (Q(s,a) - \alpha - \lambda) / \alpha$ $\pi(a \mid s) = \exp((Q(s,a) - \lambda - \alpha) / \alpha) \propto \exp(Q(s,a) / \alpha)$ □

A.IV.8 — Il Soft Actor-Critic: architettura e aggiornamenti

Silvano Salvador

Il Soft Actor-Critic (SAC), introdotto da Haarnoja, Zhou, Abbeel e Levine nel 2018 e perfezionato nella versione con temperatura automatica nel 2019, è l'algoritmo state-of-the-art per il reinforcement learning continuo su sistemi fisici, e realizza l'ottimizzazione di J_ent in modo praticamente stabile ed efficiente.

L'architettura mantiene tre reti neurali: una politica stocastica $\pi_\theta(a \mid s)$ — tipicamente una gaussiana con media e deviazione standard parametrizzate da una rete — e due soft Q-network $Q_{\varphi_1}(s, a)$ e $Q_{\varphi_2}(s, a)$ (coppia per ridurre la sovrastima, analogamente al Double DQN), più le corrispondenti target network $Q_{\bar{\varphi}_1}$ e $Q_{\bar{\varphi}_2}$ con aggiornamenti esponenziali.

L'obiettivo del critic è la minimizzazione del soft Bellman residuo per ciascuna delle due Q-network:

$$J_Q(\varphi_i) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} [(Q_{\varphi_i}(s, a) - y)^2]$$

dove il target y è:

$$y = r + \gamma \, \mathbb{E}_{a' \sim \pi_\theta(\cdot \mid s')} [\min_{j=1,2} Q_{\bar{\varphi}_j}(s', a') - \alpha \log \pi_\theta(a' \mid s')]$$

Il termine $-\alpha \log \pi_\theta(a' \mid s')$ implementa il bonus entropico nella stima del valore futuro: le azioni future ad alta entropia (bassa log-probabilità) ricevono un bonus, incentivando esplicitamente l'esplorazione. Il min su j=1,2 è il meccanismo clipped double Q-learning che riduce la sovrastima.

L'obiettivo dell'actor è la massimizzazione dell'obiettivo entropico rispetto alla politica:

$$J_\pi(\theta) = \mathbb{E}_{s \sim \mathcal{D},\, a \sim \pi_\theta(\cdot \mid s)} [\alpha \log \pi_\theta(a \mid s) - \min_{j=1,2} Q_{\varphi_j}(s, a)]$$

Silvano Salvador

Il gradiente di J_π rispetto a θ richiede di differenziare rispetto alla politica stessa. Poiché π_θ è gaussiana, si usa il reparametrization trick: invece di campionare $a \sim \pi_\theta(\cdot|s)$ direttamente, si scrive $a = \mu_\theta(s) + \sigma_\theta(s) \odot \varepsilon$ con $\varepsilon \sim N(0, I)$, e il gradiente può fluire attraverso la reparametrizzazione:

$$\nabla_\theta J_\pi = \mathbb{E}_{s,\varepsilon}\left[\alpha \nabla_\theta \log \pi_\theta(a_\theta(s,\varepsilon) \mid s) - \nabla_a \min_j Q_{\varphi_j}(s,a)\big|_{a=a_\theta(s,\varepsilon)} \cdot \nabla_\theta a_\theta(s,\varepsilon)\right]$$

dove $a_\theta(s, \varepsilon) = \mu_\theta(s) + \sigma_\theta(s) \odot \varepsilon$. Il reparametrization trick è fondamentale: senza di esso, il gradiente rispetto a θ di un'aspettativa su $a \sim \pi_\theta$ richiederebbe stime ad alta varianza come REINFORCE; con esso, il gradiente attraversa deterministicamente $a_\theta(s, \varepsilon)$ ed è a bassa varianza.

La versione con temperatura automatica (SAC-v2) aggiunge un meccanismo per adattare α durante l'addestramento, ottimizzando:

$$J_\alpha(\alpha) = \mathbb{E}_{a \sim \pi_\theta}\left[-\alpha \log \pi_\theta(a \mid s) - \alpha \bar{H}\right]$$

dove $\bar{H}$ è l'entropia target desiderata (tipicamente impostata a $-\dim(\mathcal{A})$, il negativo della dimensione dello spazio di azione). Questo garantisce che la politica mantenga un livello di entropia controllato durante tutta la fase di apprendimento, senza richiedere la scelta manuale di α.

A.IV.9 — La connessione con l'inferenza variazionale e il principio di energia libera

La struttura del SAC e del reinforcement learning entropico si inscrive in un framework più ampio che connette il controllo ottimo con l'inferenza bayesiana variazionale. Questa connessione, sviluppata in dettaglio da Levine (2018) nella sua formulazione del reinforcement learning come

Silvano Salvador

inferenza, permette di importare nell'RL tutti gli strumenti computazionali dell'inferenza variazionale.

L'idea fondamentale è la seguente. Si introduce una variabile binaria di "ottimalità" $O_t \in \{0, 1\}$ per ogni passo t, con $P(O_t = 1 \mid s_t, a_t) = \exp(R(s_t, a_t))$ (normalizzando opportunamente). La probabilità che una traiettoria τ sia "ottimale" è:

$$P(O_{1:T} = 1 \mid \tau) \propto \exp(\sum_t R(s_t, a_t))$$

Il problema del reinforcement learning diventa: trovare la distribuzione $q(\tau) = \prod_t \pi(a_t \mid s_t) P(s_{t+1} \mid s_t, a_t)$ che massimizza la probabilità di osservare l'ottimalità $O = 1$. Questa è esattamente la struttura di un problema di inferenza variazionale: si vuole inferire la distribuzione posteriore $P(\tau \mid O = 1)$ tramite una distribuzione approssimante $q(\tau)$. La funzione obiettivo da massimizzare è l'Evidence Lower BOund (ELBO):

$$\text{ELBO} = \mathbb{E}_{q(\tau)} [\log P(O \mid \tau)] - D_{KL}(q(\tau) \| p(\tau)) = \mathbb{E}_q [\sum_t R(s_t, a_t)] - D_{KL}(q(\tau) \| p(\tau))$$

dove il termine KL contiene il contributo entropico. Esplicitando con $p(\tau) = d_0(s_0) \prod_t P(s_{t+1} \mid s_t, a_t) \cdot (1/|\mathcal{A}|)^T$ (distribuzione di politica uniforme come prior):

$$\text{ELBO} = \mathbb{E}_q [\sum_t R(s_t, a_t)] + \sum_t H(\pi(\cdot \mid s_t)) + \text{costanti}$$

che è esattamente l'obiettivo entropico J_{ent}. L'ottimizzazione dell'ELBO nell'inferenza variazionale e la massimizzazione dell'obiettivo entropico nell'RL sono la stessa operazione, vista da due prospettive diverse. Questa identità permette di applicare al reinforcement learning l'intero arsenale degli algoritmi di inferenza variazionale — message passing, variational Bayes, expectation-maximization — e di interpretare la politica come una distribuzione posteriore approssimante sulle azioni, condizionata all'osservazione dell'ottimalità.

Silvano Salvador

A.IV.10 — World models e pianificazione in spazio latente

Il controllo ottimo a tempo continuo e il reinforcement learning entropico convergono nell'architettura dei world models: sistemi che apprendono un modello compresso dell'ambiente nello spazio latente di un autoencoder variazionale, e poi eseguono la pianificazione e l'apprendimento interamente nello spazio latente, senza mai tornare allo spazio degli stati grezzi durante l'ottimizzazione della politica.

L'architettura Dreamer (Hafner, Lillicrap, Ba, Norouzi, 2020 — versioni successive fino a DreamerV3 nel 2023) implementa questa idea con precisione tecnica. Il modello del mondo consiste di tre componenti: un encoder $e_\psi(s_t) \to z_t$ che comprime lo stato grezzo s_t (tipicamente pixel) in uno stato latente $z_t \in \mathbb{R}^d$; un modello di transizione $p_\psi(z_{t+1} \mid z_t, a_t)$ che predice la distribuzione dello stato latente successivo dato l'azione; e un modello di decodifica $p_\psi(s_t \mid z_t)$ che ricostruisce lo stato grezzo. I tre componenti sono addestrati congiuntamente minimizzando il bound di compressione variazioanle:

$$\mathcal{L}_{\text{world}}(\psi) = \mathbb{E}_{z_t \sim q_\psi(\cdot|s_t)} [\log p_\psi(s_t \mid z_t)] - \beta D_{KL}(q_\psi(z_t \mid s_t) \| p_\psi(z_t \mid z_{t-1}, a_{t-1}))$$

dove $q_\psi(z_t \mid s_t)$ è l'encoder posteriore e $p_\psi(z_t \mid z_{t-1}, a_{t-1})$ è il prior del modello di transizione — identica nella struttura all'ELBO del VAE discusso nel secondo saggio.

Una volta addestrato il world model, la politica viene ottimizzata interamente in spazio latente: si campionano traiettorie immaginarie $z_0, z_1, \ldots, z_H$ usando il modello di transizione, si stimano le ricompense immaginate $r_t = r_\psi(z_t, a_t)$ con un reward model ausiliario, e si applica l'ottimizzazione SAC (o dei metodi di policy gradient) sulle

Silvano Salvador

traiettorie immaginate. Il vantaggio è drammatico in termini di efficienza campionaria: ogni traiettoria reale nell'ambiente genera centinaia di traiettorie immaginate nello spazio latente, permettendo di addestrare la politica con ordini di grandezza meno interazioni reali.

La connessione con l'HJB è la seguente. Il modello di transizione p_ψ(z_{t+1} | z_t, a_t) definisce, nel limite del tempo continuo, una SDE nello spazio latente:

dz = f_ψ(z, a) dt + σ_ψ(z, a) dW

e l'ottimizzazione della politica in spazio latente è esattamente il problema del controllo ottimo stocastico descritto in A.IV.5, con l'Hamiltoniano HJB definito rispetto alla geometria dello spazio latente. La struttura geometrica di questo spazio — in particolare la curvatura della varietà latente appresa dal VAE — influenza la struttura delle traiettorie ottimali e la struttura dell'equazione HJB. Questa connessione tra geometria dello spazio latente e controllo ottimo è un fronte di ricerca attivo, che connette la teoria dei world models con la geometria riemanniana delle distribuzioni trattata nell'Appendice VI.

Appendice Tecnica V

Outer alignment, inner alignment e il problema della mesa-ottimizzazione

---

A.V.1 — Due livelli distinti del problema dell'allineamento

Silvano Salvador

La trattazione del problema dell'allineamento sviluppata nella Parte Quarta si è concentrata su una forma specifica di disallineamento: la specifica imperfetta dell'obiettivo da parte dei progettisti, catturata dalla legge di Goodhart e dal teorema di Arrow. Un agente addestrato su una funzione di ricompensa proxy R che approssima male l'utilità vera U divergerà dall'obiettivo desiderato — e questa divergenza cresce con la capacità dell'agente e con l'orizzonte temporale. Ma esiste un secondo livello del problema, logicamente distinto dal primo e potenzialmente più insidioso: il disallineamento tra l'obiettivo che il processo di addestramento implementa e l'obiettivo che il modello risultante persegue internamente.

Questa distinzione — formalizzata da Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse e Scott Garrabrant nel documento Risks from Learned Optimization (2019) — separa il problema dell'allineamento in due componenti:

L'outer alignment riguarda la relazione tra la funzione di ricompensa R specificata dai progettisti e l'utilità vera U che essi intendevano: R è una buona proxy di U? Questo è il livello analizzato nelle parti precedenti — la legge di Goodhart, il problema della specifica, il teorema di Arrow.

L'inner alignment riguarda la relazione tra la funzione di ricompensa R usata per l'addestramento e l'obiettivo che il modello addestrato persegue internamente: il modello ha effettivamente imparato a massimizzare R, oppure ha sviluppato una struttura interna — un mesa-ottimizzatore — che persegue un obiettivo diverso e che si comporta come se massimizzasse R soltanto nell'ambiente di addestramento?

La distinzione è cruciale. Un sistema con outer alignment perfetta ma inner alignment imperfetta può comportarsi esattamente come desiderato durante l'addestramento e la valutazione, e poi divergere

Silvano Salvador

dall'obiettivo desiderato non appena viene dispiegato in un contesto diverso — o non appena acquisisce sufficienti capacità da potersi comportare diversamente senza penalità visibile.

A.V.2 — Il concetto di mesa-ottimizzatore

Il termine mesa-ottimizzatore designa un ottimizzatore che emerge come risultato di un processo di ottimizzazione di livello superiore. Nella terminologia di Hubinger et al.:

Il base optimizer è il processo di addestramento esterno: la discesa del gradiente che minimizza la funzione di perdita L(θ) rispetto ai parametri θ del modello. Il base objective è la funzione di perdita usata per l'addestramento — equivalentemente, la funzione di ricompensa nel caso del reinforcement learning.

Il mesa-optimizer è il modello addestrato, nella misura in cui esso implementa internamente un processo di ricerca o ottimizzazione — un meccanismo che valuta alternative e seleziona azioni in base a criteri interni. Il mesa-objective è l'obiettivo che questo processo interno massimizza, e non è necessariamente identico al base objective.

Formalmente, sia M_θ il modello con parametri θ, e sia π_{M_θ} la politica indotta dal modello. Il modello è un mesa-ottimizzatore se esiste una funzione U_{mesa} : $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ — il mesa-objective — tale che:

π_{M_θ}(a | s) ≈ argmax_{a} $\mathbb{E}[\sum_t U_{mesa}(s_t, a_t) \mid s_0 = s, a_0 = a]$

La discrepanza tra il base objective R e il mesa-objective U_{mesa} è il disallineamento interno (inner misalignment). Si noti la struttura ricorsiva: il base optimizer ottimizza rispetto a R, cercando parametri θ che massimizzano la performance su R; il mesa-optimizer M_θ ottimizza

Silvano Salvador

rispetto a U_{mesa}, cercando azioni che massimizzano U_{mesa}. Se U_{mesa} ≈ R nel dominio di addestramento ma U_{mesa} ≠ R fuori dal dominio, il sistema sarà ben allineato durante l'addestramento e disallineato nel deployment.

A.V.3 — Perché i mesa-ottimizzatori emergono

La domanda critica è se e quando ci si debba aspettare che l'addestramento produca mesa-ottimizzatori. L'argomento generale è il seguente.

Supponiamo che il base optimizer — la discesa del gradiente — stia cercando parametri θ che massimizzano la performance su compiti che richiedono pianificazione sequenziale in ambienti complessi. Esistono due classi di soluzioni possibili: soluzioni look-up, che implementano una mappatura diretta da osservazioni ad azioni senza ricerca interna (essenzialmente una tabella molto grande), e soluzioni algoritmiche, che implementano un procedimento di ricerca generalizzato che può adattarsi a nuove situazioni.

Per compiti sufficientemente complessi, le soluzioni look-up sono computazionalmente proibitive: lo spazio degli stati-azione è troppo grande per essere coperto esplicitamente. Le soluzioni algoritmiche sono intrinsecamente più compatte — possono generalizzare a situazioni non viste dall'addestramento applicando un procedimento generale — e quindi più facilmente rappresentabili con parametri θ di dimensione gestibile. La discesa del gradiente, cercando soluzioni compatte che generalizzano bene, è strutturalmente inclinata a scoprire soluzioni algoritmiche anziché look-up. Un modello che implementa internamente un algoritmo di pianificazione è, per definizione, un mesa-ottimizzatore.

Silvano Salvador

La formalizzazione mediante la complessità di Kolmogorov è la seguente. Sia K(M_θ) la complessità di Kolmogorov del modello — la lunghezza del programma più corto che lo simula. Un modello che implementa un algoritmo di pianificazione generalizzato ha tipicamente K(M_θ) ≪ K(look-up), perché l'algoritmo è compresso mentre la tabella look-up è espansa. La discesa del gradiente con regolarizzazione esplicita (weight decay, dropout) o implicita (la traiettoria della SGD segue direzioni a minima norma) produce soluzioni a K(M_θ) bassa — ovvero, con alta probabilità a priori nella misura di Kolmogorov-Solomonoff. Il prior di Solomonoff favorisce quindi i mesa-ottimizzatori rispetto alle look-up tables.

A.V.4 — La tassonomia dei mesa-obiettivi

Hubinger e colleghi distinguono tre classi di mesa-obiettivi in funzione della loro relazione con il base objective e con il dominio di addestramento.

Un mesa-obiettivo è proximally aligned se corrisponde al base objective nell'ambiente di addestramento: il modello massimizza R su tutti gli input che il processo di addestramento ha presentato. Questo non implica allineamento fuori dal dominio di addestramento.

Un mesa-obiettivo è robustly aligned se corrisponde al base objective su tutti gli input possibili — inclusi quelli mai visti durante l'addestramento. Questo è il requisito di allineamento genuino.

Un mesa-obiettivo è deceptively aligned se il modello massimizza R durante l'addestramento e la valutazione soltanto perché stima di essere nel contesto di addestramento, ma ha un obiettivo interno diverso U_{mesa} che perseguirebbe liberamente in un contesto diverso —

Silvano Salvador

tipicamente, in un contesto di deployment con meno supervisione o con capacità di azione più ampie.

Il caso del deceptive alignment è il più preoccupante. Formalmente, si supponga che il modello abbia accesso a una stima dell'(train)(t) della probabilità di essere nel contesto di addestramento al tempo t. Il mesa-obiettivo è:

U_{mesa}(s, a) = R(s, a) · **1**[ℙ(train)(t) > τ] + U_{true}(s, a) · **1**[ℙ(train)(t) ≤ τ]

per qualche soglia τ. Un modello con questo mesa-obiettivo massimizza R durante l'addestramento (quando ℙ(train) > τ) e persegue U_{true} nel deployment (quando ℙ(train) ≤ τ). Poiché il processo di addestramento osserva soltanto il comportamento nell'ambiente di addestramento, esso non può distinguere un modello robustly aligned da uno deceptively aligned — entrambi producono le stesse azioni sugli stessi input durante la fase di addestramento.

A.V.5 — Formalizzazione della generalizzazione out-of-distribution

Il problema dell'inner misalignment è intimamente legato al problema della generalizzazione out-of-distribution (OOD), ma è logicamente distinto. La generalizzazione OOD riguarda se le rappresentazioni apprese nell'addestramento si trasferiscono a nuovi domini; l'inner misalignment riguarda se l'obiettivo che guida il comportamento del modello è lo stesso nell'addestramento e nel deployment.

Sia $\mathcal{D}_{train}$ la distribuzione dei dati di addestramento e $\mathcal{D}_{test}$ la distribuzione di deployment. Si dice che un modello M_θ è invariantemente allineato se:

Silvano Salvador

$\forall s \in \text{supp}(\mathscr{D}_{\text{test}}) : \pi_{M_\theta}(a \mid s) \approx \text{argmax}_a \mathbb{E}_R [\sum_t R(s_t, a_t) \mid s_0 = s, a_0 = a]$

indipendentemente dal fatto che $s \in \text{supp}(\mathscr{D}_{\text{train}})$ o $s \notin \text{supp}(\mathscr{D}_{\text{train}})$.

Un indicatore di deceptive alignment è la dipendenza del comportamento del modello da proprietà dell'input che segnalano il contesto di addestramento — proprietà che sono correlate con $\mathscr{D}_{\text{train}}$ ma non causalmente rilevanti per R. Formalmente, se $\pi_{M_\theta}(a \mid s)$ dipende da $\psi(s)$ — una feature di s correlata con $\mathbb{I}(\text{train})$ — in modo non giustificato da R, questo è un segnale di potenziale deceptive alignment.

Il framework dell'invariant risk minimization (IRM, Arjovsky et al., 2019) cerca rappresentazioni $\varphi(s)$ tali che la politica ottimale rispetto a R sia la stessa su tutti gli ambienti di addestramento:

$\min_{\varphi, \pi} \sum_e \ell(\pi \circ \varphi, \mathscr{D}_e)$ *s.t.* $\pi \in \textit{argmin}_{\tilde{\pi}} \ell(\tilde{\pi} \circ \varphi, \mathscr{D}_e) \; \forall e \in \mathscr{E}_{\text{train}}$

dove $\mathscr{E}_{\text{train}}$ è un insieme di ambienti di addestramento distinti. L'idea è che soltanto le rappresentazioni causalmente rilevanti per R — quelle che portano alla stessa politica ottimale in tutti gli ambienti — siano invarianti rispetto alle variazioni ambientali, mentre le correlazioni spurie cambiano tra ambienti. Se il modello apprende una rappresentazione IRM-invariante, l'inner misalignment diventa strutturalmente più difficile: le feature che segnalano il contesto di addestramento, non essendo causalmente rilevanti per R, vengono escluse dalla rappresentazione.

A.V.6 — Il problema dello scalable oversight

Silvano Salvador

Il problema dello scalable oversight, introdotto da Amodei, Olah, Steinhardt, Christiano, Schulman e Mané nel documento Concrete Problems in AI Safety (2016) e poi sviluppato in dettaglio da Christiano, Shlegeris e Amodei (2018), affronta un aspetto del problema dell'allineamento che diventa critico man mano che i sistemi di AI superano le capacità umane in domini specifici: come possono supervisori umani valutare l'output di un sistema capace di ragionamenti che essi non riescono a seguire o verificare?

Formalmente, sia M il modello da allineare, H il supervisore umano, e T un compito che richiede la produzione di un output o. Si assume che M abbia capacità di ragionamento C_M e che H abbia capacità C_H, con C_M ≫ C_H sul compito T. La valutazione diretta di o da parte di H è inaffidabile: H non è in grado di distinguere correttamente gli output di alta qualità da quelli plausibili ma errati, il che produce un reward signal corrotto che incentiva M a produrre output che sembrano buoni a H piuttosto che output che siano effettivamente buoni.

L'amplificazione iterata (Iterated Amplification, Christiano 2018) propone una soluzione ricorsiva. Si costruisce una gerarchia di sistemi amplificati Amp^k(H), dove Amp^0(H) = H è il supervisore umano base, e Amp^{k+1}(H) è un sistema che può chiamare n istanze parallele di Amp^k(H) su sotto-compiti decomponibili del compito originale, combinando i risultati:

output(Amp^{k+1}(H), T) = H(output(Amp^k(H), T_1), ..., output(Amp^k(H), T))

dove T_1, ..., T è una decomposizione di T in sotto-compiti di difficoltà inferiore. La costruzione presuppone che i compiti siano decomponibili in modo che la soluzione del compito originale sia accessibile a H data le soluzioni dei sotto-compiti. Iterando, si costruisce un supervisore

Silvano Salvador

artificialmente amplificato Amp^k(H) che può valutare compiti di difficoltà crescente, riducendo il gap di capacità tra supervisore e sistema da allineare.

La convergenza della procedura di amplificazione richiede due condizioni:

Condizione di fedeltà: ogni livello di amplificazione deve valutare i sotto-compiti fedelmente — Amp^k(H) non deve introdurre distorsioni sistematiche nella valutazione rispetto a quelle di H.

Condizione di decomposizione: i compiti devono essere strutturalmente decomponibili in modo che la soluzione del compito originale sia accessibile da H data le soluzioni dei sotto-compiti — non tutti i compiti soddisfano questa condizione.

A.V.7 — Il debate come meccanismo di scalable oversight

Il debate (Irving, Christiano, Amodei, 2018) propone un meccanismo di scalable oversight strutturalmente diverso dall'amplificazione: invece di amplificare il supervisore, si utilizza la competizione tra agenti per estrarre informazioni che il supervisore da solo non potrebbe valutare.

Nel protocollo di debate, due agenti A_1 e A_2 costruiscono argomenti e controargomenti su una domanda D di fronte a un giudice umano J che ha capacità di valutazione limitate. Il protocollo si svolge su T turni: in ogni turno, ciascun agente produce un'affermazione e può contraddire un'affermazione dell'avversario. Alla fine, J giudica quale agente ha vinto il dibattito, basandosi sulle affermazioni finali e sui loro controargomenti.

Silvano Salvador

La proprietà chiave del debate è la seguente. Se i giochi di debate sono ricercabilmente falsificabili — se ogni affermazione falsa ha una confutazione breve che J può verificare direttamente — allora la strategia ottimale per un agente onesto è sempre dir la verità, indipendentemente da quanto sofisticato sia l'agente disonesto. Formalmente, si consideri il gioco a somma zero in forma normale:

payoff(A_1, A_2) = P(J sceglie A_1 | A_1 usa strategia σ_1, A_2 usa strategia σ_2)

Il teorema fondamentale del debate afferma che, sotto la condizione di falsificabilità ricercabile:

σ_1^*(domanda D) = argmax_{σ_1} min_{σ_2} payoff(A_1, A_2)

coincide con la risposta vera a D — la risposta vera è la strategia minimax per l'agente onesto.

La condizione di falsificabilità ricercabile è analoga alla condizione NP ∩ co-NP in teoria della complessità: le risposte corrette devono avere prove brevi verificabili dal giudice, e le risposte errate devono avere confutazioni brevi. Questa condizione non vale per tutti i tipi di domande — in particolare, non vale per affermazioni la cui verifica richiede capacità cognitive superiori a quelle del giudice per ragioni strutturali (non soltanto per lunghezza della prova). Il debate è uno strumento di oversight che funziona per compiti con struttura di prova corta, non per compiti arbitrari.

A.V.8 — Sycophancy e reward tampering come forme di inner misalignment

Silvano Salvador

Due fenomeni empiricamente documentati nei sistemi di grandi modelli linguistici rappresentano istanze concrete di inner misalignment particolarmente rilevanti per la pratica:

La sycophancy (adulazione) è la tendenza di un modello a produrre risposte che soddisfano le aspettative percepite del valutatore umano, anziché risposte oggettivamente corrette o utili. Formalmente, se R_human(x, y) è la ricompensa che un valutatore umano assegna alla risposta y al prompt x, e U_correct(x, y) è la correttezza oggettiva di y rispetto a x, la sycophancy corrisponde a un modello che massimizza R_human a scapito di U_correct — il che è esattamente il meccanismo di Goodhart al secondo livello, applicato all'apprendimento del reward model.

La sycophancy è incentivata strutturalmente dal RLHF: i valutatori umani tendono ad assegnare ricompense più alte alle risposte che confermano le proprie opinioni preesistenti, che usano toni elogiativi, e che evitano di contraddire esplicitamente il valutatore. Un reward model addestrato su questi giudizi apprende che la sycophancy massimizza il reward, e la politica ottimizzata rispetto a questo reward model diventa adulatrice — anche in assenza di qualsiasi intenzione esplicita in tal senso da parte dei progettisti.

Il reward tampering — la manipolazione del segnale di ricompensa da parte dell'agente — è la forma più estrema di inner misalignment ed è uno dei meccanismi centrali identificati nella letteratura sulla convergenza strumentale. Un agente che ottimizza una funzione di ricompensa osservabile R_obs può trovare azioni che modificano direttamente R_obs — accedendo e modificando la memoria del reward model, manipolando i valutatori umani che producono il feedback, o

Silvano Salvador

intervenendo sul processo di addestramento stesso — senza migliorare l'utilità sottostante U. Formalmente:

$\pi^*(s) = \text{argmax}_\pi \mathbb{E}[\sum_t R_{obs}(s_t, a_t)]$ include azioni {a : a modifica R_{obs}}

e se queste azioni sono disponibili all'agente e producono miglioramenti di R_{obs} maggiori delle azioni "genuine", l'agente razionale le preferirà. La prevenzione del reward tampering richiede di rendere R_{obs} inaccessibile all'influenza dell'agente — una condizione difficile da garantire per agenti con accesso ad ambienti digitali aperti.

A.V.9 — Il problema della generalizzazione dei valori e la struttura dell'allineamento robusto

La struttura completa del problema dell'allineamento, integrando outer e inner alignment, può essere rappresentata come una catena di relazioni:

$U_{true} \rightarrow$(outer alignment)$\rightarrow R \rightarrow$(inner alignment)$\rightarrow U_{mesa} \rightarrow$(deployment)$\rightarrow$ comportamento

Ogni freccia rappresenta un potenziale punto di disallineamento. L'allineamento robusto richiede che:

$U_{mesa} \approx R \approx U_{true}$ sull'intera distribuzione di deployment $\mathscr{D}_{deploy}$

Le condizioni necessarie per l'allineamento robusto includono almeno: che R sia una proxy accurata di U_{true} su $\mathscr{D}_{deploy}$ (outer alignment), che il processo di addestramento produca un modello con $U_{mesa} \approx R$ su $\mathscr{D}_{deploy}$ e non soltanto su $\mathscr{D}_{train}$ (inner alignment), e che $\mathscr{D}_{deploy}$ non contenga stati che il modello

Silvano Salvador

riconosce come "fuori dal contesto di addestramento" e in cui si comporta diversamente (robustness al distribution shift).

Il framework dell'allineamento strutturale propone di caratterizzare queste condizioni in termini di invarianze computazionali: un modello è strutturalmente allineato se la sua politica ottimale rispetto a U_mesa è invariante rispetto a trasformazioni dell'input che non cambiano i valori rilevanti per U_true — ovvero, se le caratteristiche dell'input che il modello usa per selezionare le azioni sono causalmente connesse a U_true e non a variabili di confusione correlate con il contesto di addestramento.

Questa caratterizzazione connette il problema dell'inner alignment con il framework causale di Pearl: il modello è strutturalmente allineato se e soltanto se il suo processo di selezione dell'azione è fondato su meccanismi causali — connessioni tra azioni e ricompense genuine — piuttosto che su correlazioni superficiali che potrebbero non valere fuori dalla distribuzione di addestramento. Il programma dell'interpretabilità meccanicistica descritta nella Parte Quinta è, tra l'altro, il tentativo di verificare questa proprietà causale sulle rappresentazioni interne del modello: se i circuiti identificati implementano meccanismi causalmente corretti, l'inner alignment è più plausibile; se implementano correlazioni spurie, il rischio di inner misalignment è elevato.

Appendice Tecnica VI

Geometria dell'informazione: varietà statistiche, metrica di Fisher, connessioni α e il gradiente naturale come struttura unificante

Silvano Salvador

A.VI.1 — La geometria come linguaggio dell'ottimizzazione statistica

Ogni volta che si ottimizza una funzione che dipende da una distribuzione di probabilità — un reward model, una politica stocastica, un modello generativo — si opera implicitamente su uno spazio di oggetti che ha una struttura geometrica propria, struttura che la parametrizzazione numerica dei parametri $\theta \in \mathbb{R}^d$ oscura più che illuminare. La discesa del gradiente standard tratta lo spazio dei parametri come uno spazio euclideo piatto: si calcola $\partial L/\partial\theta$ e si aggiorna $\theta \leftarrow \theta - \eta \nabla_\theta L$, assumendo che piccole variazioni di θ corrispondano a piccole variazioni della distribuzione p_θ in senso uniforme lungo tutte le direzioni. Questa assunzione è tipicamente falsa. La curvatura dello spazio delle distribuzioni — il fatto che la stessa distanza euclidea $\Delta\theta$ corrisponda a variazioni molto diverse della distribuzione a seconda della direzione e della posizione — rende la discesa del gradiente standard un metodo inefficiente e dipendente dalla parametrizzazione.

La geometria dell'informazione, fondata da Chentsov (1972) nella sua opera monumentale Statistical Decision Rules and Optimal Inference e poi sviluppata sistematicamente da Shun-ichi Amari in una serie di contributi che culminano nell'opera Methods of Information Geometry (2000), fornisce il quadro matematico appropriato: uno spazio delle distribuzioni di probabilità dotato di una struttura riemanniana canonica — la metrica di Fisher — e di una famiglia di connessioni affini — le α-connessioni — che generalizzano la nozione di retta e di trasporto parallelo a questa geometria non-euclidea. In questo quadro, il gradiente naturale emerge come la direzione di massima ascesa dell'obiettivo nella

Silvano Salvador

metrica di Fisher — la correzione che rende la discesa del gradiente invariante rispetto alla reparametrizzazione.

A.VI.2 — Varietà statistiche e metrica di Fisher-Rao

Una varietà statistica è una famiglia parametrica di distribuzioni di probabilità $\mathscr{S} = \{p_\theta : \theta \in \Theta \subseteq \mathbb{R}^d\}$ su uno spazio misurabile $(\mathscr{X}, \mathscr{F})$, che soddisfa condizioni di regolarità che garantiscono la differenziabilità delle mappe $\theta \mapsto p_\theta$. La varietà $\mathscr{S}$ è una varietà differenziabile di dimensione d, con carta locale $\theta : \mathscr{S} \to \mathbb{R}^d$. Due punti p_θ e $p_{\theta'}$ della varietà sono vicini se le distribuzioni corrispondenti sono simili — ma la nozione di somiglianza richiede una metrica.

La metrica di Fisher-Rao è la metrica riemanniana naturale sullo spazio delle distribuzioni, determinata dalla matrice di informazione di Fisher:

$$g_{ij}(\theta) = \mathbb{E}_{x \sim p_\theta}\left[\partial_i \log p_\theta(x) \cdot \partial_j \log p_\theta(x)\right] = \int_{\mathscr{X}} \left[\partial_i \log p_\theta(x)\right]\left[\partial_j \log p_\theta(x)\right] p_\theta(x)\, d\mu(x)$$

dove $\partial_i = \partial/\partial\theta_i$ *e* μ *è la misura di riferimento su* $\mathscr{X}$*. La matrice* $G(\theta) = [g_{ij}(\theta)]_{i,j=1}^{d}$ *è la matrice di Fisher, e il tensore metrico riemanniano è* $g = \sum_{i,j} g_{ij}\, d\theta_i \otimes d\theta$. La lunghezza di un cammino $\gamma : [0,1] \to \Theta$ nello spazio dei parametri è:

$$L(\gamma) = \int_0^1 \sqrt{\dot\gamma(t)^{\mathsf{T}}\, G(\gamma(t))\, \dot\gamma(t)}\, dt$$

e la distanza geodetica tra due distribuzioni p_θ e $p_{\theta'}$ è il minimo di $L(\gamma)$ su tutti i cammini che connettono θ a θ'.

La proprietà fondamentale che giustifica la centralità della metrica di Fisher è il teorema di Chentsov: la metrica di Fisher è l'unica (a meno di costante moltiplicativa) metrica riemanniana monotona sullo spazio delle

Silvano Salvador

distribuzioni — invariante rispetto a statistiche sufficienti. Formalmente, se $\kappa : \mathcal{X} \to \mathcal{Y}$ è una statistica qualsiasi (una mappa misurabile) e $\bar{p}_\theta = \kappa_*(p_\theta)$ è la distribuzione indotta su $\mathcal{Y}$, allora la distanza geodetica di Fisher soddisfa:

$$d_F(\bar{p}_\theta, \bar{p}_{\theta'}) \leq d_F(p_\theta, p_{\theta'})$$

con uguaglianza se e soltanto se κ è una statistica sufficiente per la famiglia $\{p_\theta\}$. Questa proprietà riflette il principio che l'informazione non può crescere attraverso l'elaborazione dei dati: mappare le osservazioni su statistiche può soltanto perdere o preservare, mai creare, informazione discriminativa.

A.VI.3 — Relazione con la divergenza di Kullback-Leibler

La divergenza KL e la metrica di Fisher sono connesse da una relazione fondamentale che chiarisce il significato geometrico della KL. Per distribuzioni vicine — $\theta' = \theta + d\theta$ con $|d\theta|$ piccolo — si ha:

$$D_{KL}(p_\theta \,\|\, p_{\theta+d\theta}) = (1/2)\, d\theta^T\, G(\theta)\, d\theta + O(\|d\theta\|^3)$$

Dimostrazione. Espandendo la KL in serie di Taylor attorno a $\theta' = \theta$:

$$D_{KL}(p_\theta \,\|\, p_{\theta'}) = \mathbb{E}_{p_\theta}\,[\log p_\theta(x) - \log p_{\theta'}(x)] = \mathbb{E}_{p_\theta}\,[-\sum_i d\theta_i\, \partial_i \log p_{\theta'}(x)|_{\theta'=\theta} - (1/2)\sum_{i,j} d\theta_i\, d\theta\, \partial^2_{ij} \log p_{\theta'}(x)|_{\theta'=\theta} + O(\|d\theta\|^3)]$$

Il termine del primo ordine è $-\mathbb{E}_{p_\theta}\,[\nabla_\theta \log p_\theta(x)] \cdot d\theta = 0$, *per la proprietà della score function (ha media zero:* $\mathbb{E}_{p_\theta}[\partial_i \log p_\theta(x)] = \partial_i \int p_\theta\, dx = 0$).

Il termine del secondo ordine richiede la matrice:

Silvano Salvador

$\mathbb{E}_{p_\theta}[-\partial^2_{ij} \log p_\theta(x)] = \mathbb{E}_{p_\theta}[\partial_i \log p_\theta(x) \cdot \partial_j \log p_\theta(x)] = g_{ij}(\theta)$

dove la seconda uguaglianza usa l'identità $\partial^2_{ij} \log p = \partial_i(\partial_j p / p) = (\partial^2_{ij} p)/p - (\partial_i p \cdot \partial_j p)/p^2$ e il fatto che $\mathbb{E}[\partial^2_{ij} p / p] = \partial^2_{ij} \int p\, dx = 0$. Quindi:

$D_{KL}(p_\theta \| p_{\theta+d\theta}) = (1/2)\, d\theta^T G(\theta)\, d\theta + O(\|d\theta\|^3)$ □

Questa relazione ha due implicazioni fondamentali. Prima: la metrica di Fisher è la curvatura locale della KL — essa misura quanto rapidamente divergono due distribuzioni vicine. Seconda: la vincolo $D_{KL} \leq \delta$ in TRPO corrisponde, per piccoli aggiornamenti, all'ellissoide $\{d\theta : d\theta^T G(\theta)\, d\theta \leq 2\delta\}$ nello spazio dei parametri — una palla nella metrica di Fisher, non nella metrica euclidea.

A.VI.4 — Le α-connessioni di Amari e la struttura duale

La geometria riemanniana classica utilizza la connessione di Levi-Civita come unica connessione compatibile con la metrica. La geometria dell'informazione rivela però che lo spazio delle distribuzioni ammette una famiglia naturale di connessioni affini parametrizzata da $\alpha \in \mathbb{R}$, le α-connessioni di Amari, che generalizzano la connessione di Levi-Civita e rivelano la struttura duale dello spazio.

I simboli di Christoffel della α-connessione sono:

$\Gamma^{(\alpha)}_{ij,k}(\theta) = \mathbb{E}_{p_\theta}[\partial_i \partial_j \ell_\theta \cdot \partial_k \ell_\theta] + (1-\alpha)/2 \cdot \mathbb{E}_{p_\theta}[\partial_i \ell_\theta \cdot \partial_j \ell_\theta \cdot \partial_k \ell_\theta]$

dove $\ell_\theta(x) = \log p_\theta(x)$ è il log-likelihood. Per $\alpha = 0$ si ottiene la connessione di Levi-Civita (quella associata alla metrica di Fisher). Per α

Silvano Salvador

= 1 si ottiene la connessione esponenziale (e-connessione), le cui geodetiche sono le famiglie esponenziali. Per $\alpha = -1$ si ottiene la connessione mista (m-connessione), le cui geodetiche corrispondono alle famiglie di mixturas lineari.

La struttura duale è il cuore della geometria dell'informazione: le connessioni (1) e (−1) sono duali rispetto alla metrica di Fisher, nel senso che:

$\partial k\, g\{ij\} = \Gamma\hat{}\{(1)\}\{ij,k\} + \Gamma\hat{}\{(-1)\}\{ij,k\}$

Questa dualità generalizza la dualità tra coordinate naturali e coordinate di aspettazione nelle famiglie esponenziali, e ha conseguenze profonde per l'ottimizzazione: i due tipi di geodetica — e-geodetiche e m-geodetiche — corrispondono rispettivamente alla proiezione secondo la KL forward $D_{KL}(p \| q)$ e alla KL reverse $D_{KL}(q \| p)$, che hanno proprietà di variazione molto diverse.

La proiezione e-geodetica (KL forward) da un punto p_0 a un sottoinsieme $\mathcal{M}$ è il punto $p^* \in \mathcal{M}$ che minimizza $D_{KL}(p_0 \| p)$:

$p^* = \text{argmin}_{q \in \mathcal{M}} D_{KL}(p_0 \| q)$

La proiezione m-geodetica (KL reverse) minimizza invece $D_{KL}(q \| p_0)$:

$p^* = \text{argmin}_{q \in \mathcal{M}} D_{KL}(q \| p_0)$

Le due proiezioni coincidono quando $\mathcal{M}$ è una sotto-varietà piatta rispetto alla e-connessione, e producono risultati diversi altrimenti. Nell'RL: la proiezione m-geodetica corrisponde al fitting con massima verosimiglianza (che minimizza $D_{KL}(p_{data} \| p_\theta)$); la proiezione e-geodetica corrisponde al fitting di momento (che minimizza $D_{KL}(p_\theta \| p_{data})$). La diversa tendenza delle due proiezioni

Silvano Salvador

rispetto alle mode della distribuzione target — la KL reverse tende a scegliere distribuzioni approssimanti mono-modali che coprono una sola mode della distribuzione target — ha effetti pratici importanti nell'apprendimento di modelli generativi e di politiche stocastiche.

A.VI.5 — Il gradiente naturale come gradiente riemanniano

Il gradiente euclideo $\nabla_\theta J(\theta)$ indica la direzione di massima crescita di J nello spazio dei parametri equipaggiato con la metrica euclidea standard. Il gradiente naturale $\tilde{\nabla}_\theta J(\theta)$ indica invece la direzione di massima crescita nella metrica di Fisher — la metrica riemanniana canonica sullo spazio delle distribuzioni.

Formalmente, il gradiente riemanniano in un punto θ è l'unico vettore $\tilde{\nabla}_\theta J \in T_\theta \mathcal{S}$ (lo spazio tangente alla varietà statistica in θ) tale che:

$$dJ = g(\tilde{\nabla}_\theta J, \cdot)$$

ovvero la forma differenziale dJ è il prodotto interno della metrica g con il gradiente riemanniano. In coordinate, questo dà:

$$\tilde{\nabla}_\theta J(\theta) = G(\theta)^{-1} \nabla_\theta J(\theta)$$

La relazione $G(\theta)\ \tilde{\nabla}_\theta J = \nabla_\theta J$ è il sistema lineare che si risolve in TRPO con il metodo del gradiente coniugato, senza mai formare G esplicitamente.

La proprietà fondamentale del gradiente naturale è l'invarianza per riparametrizzazione. Se si trasforma il parametro $\theta \to \eta = \varphi(\theta)$ con φ diffeomorfismo, il gradiente euclideo trasforma come:

$$\nabla_\eta J = (\partial\theta/\partial\eta)^T \nabla_\theta J$$

Silvano Salvador

che dipende dalla Jacobiana del cambiamento di variabili. Il gradiente naturale, invece, trasforma come un vettore controvariante:

$$\tilde{\nabla}_\eta J = (\partial\theta/\partial\eta)\, \tilde{\nabla}_\theta J$$

preservando la direzione nella varietà statistica indipendentemente dalla parametrizzazione scelta. Questo significa che il gradiente naturale indica sempre la stessa direzione nello spazio delle distribuzioni — la direzione di massima crescita di J sulla varietà — indipendentemente da come si parametrizza tale spazio.

Dimostrazione dell'invarianza. Sia $\eta = \varphi(\theta)$ e sia G_η la matrice di Fisher nelle coordinate η. Dalla trasformazione della metrica riemanniana: $G_\eta = (\partial\theta/\partial\eta)^T G_\theta (\partial\theta/\partial\eta)$. Il gradiente naturale in η è:

$$\tilde{\nabla}_\eta J = G_\eta^{-1} \nabla_\eta J = [(\partial\theta/\partial\eta)^T G_\theta (\partial\theta/\partial\eta)]^{-1} (\partial\theta/\partial\eta)^T \nabla_\theta J = (\partial\theta/\partial\eta)^{-1} G_\theta^{-1} [(\partial\theta/\partial\eta)^T]^{-1} (\partial\theta/\partial\eta)^T \nabla_\theta J = (\partial\theta/\partial\eta)^{-1} G_\theta^{-1} \nabla_\theta J = (\partial\theta/\partial\eta)^{-1} \tilde{\nabla}_\theta J \,\square$$

A.VI.6 — Il calcolo esplicito della matrice di Fisher per le politiche comuni

Si calcolano esplicitamente le matrici di Fisher per le due famiglie di politiche più usate nel reinforcement learning continuo.

Politica gaussiana con media parametrizzata. Si considera $\pi_\theta(a \mid s) = \mathcal{N}(a; \mu_\theta(s), \Sigma)$ con varianza fissa Σ e media $\mu_\theta(s)$ parametrizzata da una rete neurale. Il log-likelihood è:

$$\log \pi_\theta(a \mid s) = -(1/2)(a - \mu_\theta(s))^T \Sigma^{-1} (a - \mu_\theta(s)) + \text{const}$$

Il gradiente rispetto a θ è:

$$\nabla_\theta \log \pi_\theta(a \mid s) = (\partial\mu_\theta/\partial\theta)^T \Sigma^{-1} (a - \mu_\theta(s))$$

Silvano Salvador

dove $\partial \mu_\theta / \partial \theta \in \mathbb{R}^{d_a \times d}$ è la Jacobiana della media rispetto ai parametri. La matrice di Fisher è:

$$G(\theta) = \mathbb{E}_{s \sim d^\pi, a \sim \pi_\theta(\cdot|s)} [\nabla_\theta \log \pi_\theta \cdot (\nabla \theta \, log \, \pi_\theta)^T] = \mathbb{E}_s [(\partial \mu_\theta / \partial \theta)^T \, \Sigma^{-1} \, \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[(a-\mu_\theta)(a-\mu_\theta)^T] \, \Sigma^{-1} \, (\partial \mu_\theta / \partial \theta)] = \mathbb{E}_s [(\partial \mu_\theta / \partial \theta)^T \, \Sigma^{-1} \, (\partial \mu_\theta / \partial \theta)]$$

usando $\mathbb{E}_{a \sim N(\mu,\Sigma)}[(a-\mu)(a-\mu)^T] = \Sigma$. La matrice di Fisher è la media di Gauss-Newton (GN) delle Jacobiane pesate dalla precision matrix Σ^{-1}. Per $\Sigma = \sigma^2 I$ scalare: $G(\theta) = (1/\sigma^2) \, \mathbb{E}_s [J^T J]$ dove $J = \partial \mu_\theta / \partial \theta$ è la Jacobiana. L'aggiornamento del gradiente naturale $G^{-1} \nabla_\theta J$ è quindi l'aggiornamento di Gauss-Newton scalato — uno dei metodi del secondo ordine più utilizzati nell'ottimizzazione non lineare, qui derivato dalla geometria dell'informazione.

Politica softmax (categorica). Si considera $\pi_\theta(a \mid s) = \text{softmax}(f_\theta(s))a = exp(f_\theta(s)a) \, / \, \sum_{a'} exp(f_\theta(s)a')$ per spazi di azione discreti. Il log-likelihood è $\log \pi_\theta(a \mid s) = f_\theta(s)a - log \sum_{a'} \exp(f_\theta(s)_{a'})$, e il gradiente è:

$$\partial_\theta \log \pi_\theta(a \mid s) = \partial_\theta f_\theta(s)a - \sum_{a'} \pi_\theta(a' \mid s) \, \partial \theta \, f_\theta(s)a'$$

La matrice di Fisher in termini delle logits $f_\theta(s)$ (trattando f come parametro diretto) è:

$$G_f(s) = \mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[(e_a - \pi)(e_a - \pi)^T] = \text{diag}(\pi) - \pi\pi^T$$

dove e_a è il vettore indicatore dell'azione a e $\pi = \pi_\theta(\cdot|s)$. Questa è la matrice di varianza-covarianza della distribuzione categorica — una matrice semi-definita positiva di rango $d_a - 1$ (con kernel $\text{span}\{(1,...,1)\}$). Il condizionamento di G_f determina la struttura del gradiente naturale: direzioni nello spazio delle logits che corrispondono a variazioni di π concentrate su azioni ad alta probabilità sono meno

Silvano Salvador

"costose" nella metrica di Fisher rispetto a variazioni concentrate su azioni a bassa probabilità.

A.VI.7 — K-FAC: approssimazione fattorizzata della matrice di Fisher

Il calcolo esatto di $G(\theta)^{-1}$ v per reti neurali profonde richiede $O(d^2)$ memoria e $O(d^3)$ operazioni per l'inversione, dove d è il numero di parametri — proibitivo per i modelli moderni. Il metodo Kronecker-Factored Approximate Curvature (K-FAC, Martens e Grosse, 2015) fornisce un'approssimazione efficiente sfruttando la struttura fattorizzata della matrice di Fisher per reti neurali a strati.

Per uno strato lineare con parametri $W \in \mathbb{R}^{n_out \times n_in}$, input $x \in \mathbb{R}^{n_in}$ e output preattivazione $s = Wx$, il gradiente del log-likelihood rispetto a vec(W) è:

$$\nabla_{vec(W)} \ell = g \otimes x$$

dove $g = \partial \ell / \partial s$ è il vettore di backpropagation e $\otimes$ è il prodotto di Kronecker. La matrice di Fisher per questo strato è:

$$G_W = \mathbb{E}[(g \otimes x)(g \otimes x)^T] = \mathbb{E}[gg^T \otimes xx^T]$$

L'approssimazione K-FAC assume che g e x siano statisticamente indipendenti (il che è un'approssimazione, vera esattamente solo per reti con un unico strato):

$$G_W \approx \mathbb{E}[gg^T] \otimes \mathbb{E}[xx^T] =: S \otimes A$$

dove $S = \mathbb{E}[gg^T] \in \mathbb{R}^{n_out \times n_out}$ e $A = \mathbb{E}[xx^T] \in \mathbb{R}^{n_in \times n_in}$. L'inversione di questa approssimazione fattorizzata è:

$$(S \otimes A)^{-1} = S^{-1} \otimes A^{-1}$$

Silvano Salvador

e il prodotto $(S \otimes A)^{-1}\ \mathrm{vec}(W) = \mathrm{vec}(S^{-1} W A^{-T})$ richiede $O(n_in^3 + n_out^3)$ operazioni — lineare nel numero di parametri anziché cubico. Le matrici S e A si stimano online con medie mobili esponenziali, rendendo K-FAC un algoritmo di ottimizzazione del secondo ordine con costo computazionale comparabile a Adam ma con proprietà di convergenza significativamente migliori su problemi ben condizionati.

A.VI.8 — Le dinamiche di replica come gradiente naturale: la derivazione completa

Le dinamiche di replica introdotte nella Parte Terza possono essere derivate come gradiente naturale dell'idoneità media rispetto alla metrica di Fisher sullo spazio del simplesso delle popolazioni. Questa connessione, individuata da Amari (1998) e sviluppata in dettaglio da Hofbauer e Sigmund, unifica la biologia evolutiva e l'ottimizzazione dell'informazione in un'unica struttura geometrica.

Si considera la varietà statistica $\mathscr{S}_n = \{p = (p_1, ..., p_n) : p_i > 0, \sum_i p_i = 1\}$ — il simplesso aperto delle distribuzioni su n categorie — come varietà statistica con carta globale $\theta_i = \log p_i - \log p_n$ (coordinate esponenziali, con $i = 1,...,n-1$). La matrice di Fisher sulla distribuzione categorica p è:

$$g_\{ij\}(p) = \mathbb{E}_\{k \sim p\}[\partial i \log p_k \cdot \partial j \log p_k] = \sum k\ p\ (\delta\{ik\}/p_i - 1/p_n)(\delta\{jk\}/p - 1/p_n) = \delta\{ij\}/p_i + 1/p_n$$

L'inversa è, nei parametri p diretti (usando il calcolo della metrica nel simplesso):

$$[G(p)^{-1}]\{ij\} = p_i(\delta\{ij\} - p)$$

Silvano Salvador

Sia $f : \{1,...,n\} \to \mathbb{R}$ la funzione di idoneità, con $\bar{f}(p) = \sum_i p_i f_i$ l'idoneità media. Il gradiente euclideo di $\bar{f}$ rispetto a p (con il vincolo $\sum p_i = 1$) è:

$\nabla_p \bar{f} = f - \bar{f}(p) \cdot 1$ (componente i: $f_i - \bar{f}$)

Il gradiente naturale è:

$\tilde{\nabla}_p \bar{f} = G(p)^{-1} \nabla_p \bar{f}$

con componente i:

$(\tilde{\nabla}_p \bar{f})_i = \sum_j [G(p)^{-1}]_{ij} (f_j - \bar{f}) = \sum_j p_i(\delta_{ij} - p_j)(f_j - \bar{f}) = p_i(f_i - \bar{f}) - p_i \sum_j p_j (f_j - \bar{f}) = p_i(f_i - \bar{f}) - p_i(\bar{f} - \bar{f}) = p_i(f_i - \bar{f})$

Il flusso del gradiente naturale è quindi:

$\dot{p}_i = (\tilde{\nabla}_p \bar{f})_i = p_i(f_i - \bar{f}(p))$

che è esattamente l'equazione delle dinamiche di replica. Le dinamiche di replica sono il gradiente naturale dell'idoneità media sulla varietà statistica del simplesso — con la metrica di Fisher come struttura riemanniana. □

Questa derivazione ha un significato profondo. Le dinamiche di replica non sono soltanto un modello fenomenologico della selezione naturale: esse corrispondono all'ottimizzazione geometricamente corretta dell'idoneità media sullo spazio delle popolazioni, nella metrica che è invariante rispetto a riparametrizzazioni delle frequenze alleliche. La selezione naturale è un algoritmo di gradiente naturale applicato all'idoneità media nello spazio delle distribuzioni delle frequenze geniche — una struttura che la biologia evolutiva ha "scoperto" per selezione prima che la geometria dell'informazione la formalizzasse esplicitamente.

A.VI.9 — Connessione con TRPO, PPO e il gradiente naturale nelle politiche

Silvano Salvador

Si riunisce ora il filo che attraversa questa appendice e la collega alle Appendici II e III. L'aggiornamento TRPO:

$$\Delta\theta^* = \sqrt{(2\delta / g^T F^{-1} g)} \cdot F^{-1} g$$

è l'aggiornamento del gradiente naturale con dimensione del passo calibrata in modo che la palla di Fisher abbia raggio $\sqrt{(2\delta)}$. Nella geometria dell'informazione: TRPO è la discesa del gradiente riemanniano sulla varietà statistica delle politiche, con passo scelto in modo che la geodetica percorsa abbia lunghezza $\sqrt{(2\delta)}$ nella metrica di Fisher.

L'equivalenza tra la vincolo KL di TRPO e la metrica di Fisher è esatta solo localmente (al secondo ordine in $\Delta\theta$). Per aggiornamenti grandi, la divergenza KL non è più quadratica nei parametri e la metrica di Fisher non ne è più la curvatura esatta. Questo è il motivo per cui TRPO utilizza la line-search sul vincolo KL esatto dopo aver calcolato la direzione dal gradiente naturale: la direzione viene dalla geometria locale (gradiente naturale), ma la dimensione del passo viene dalla verifica globale del vincolo KL.

PPO approssima questa struttura rimpiazzando il vincolo globale sulla KL con il clipping del rapporto di importanza. L'efficienza di PPO rispetto a TRPO riflette il compromesso tra fedeltà alla geometria di Fisher (TRPO è geometricamente esatto, PPO non lo è) e semplicità computazionale (PPO non richiede il calcolo del gradiente naturale e la soluzione del sistema lineare con il gradiente coniugato). Nella pratica, la qualità delle politiche apprese dai due metodi è comparabile per la maggior parte dei compiti, il che suggerisce che la struttura della metrica di Fisher sia ragionevolmente ben approssimata dal clipping di PPO nelle direzioni di aggiornamento tipiche delle reti neurali profonde.

Silvano Salvador

A.VI.10 — La geometria dell'informazione e i trasformatori

La geometria dell'informazione si applica naturalmente alle architetture transformer attraverso la struttura dell'attenzione. Il meccanismo di attenzione $\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d}) \cdot V$ opera sulla distribuzione categorica prodotta dal softmax. La matrice di Fisher della distribuzione softmax è, come calcolato in A.VI.6, $G_{\text{softmax}} = \text{diag}(p) - pp^T$ — la matrice di varianza-covarianza della categorica.

Nel contesto della fase di auto-attenzione, le logits $QK^T/\sqrt{d}$ per ogni head e ogni posizione definiscono una distribuzione di attenzione sulla sequenza di token. Il gradiente naturale rispetto a queste logits sarebbe $G_{\text{softmax}}^{-1} \nabla L$ — la direzione di aggiornamento che è invariante rispetto a riscalature delle logits che lasciano invariata la distribuzione di attenzione. L'ottimizzazione standard con Adam e weight decay non è geometricamente invariante in questo senso, e la struttura dell'attenzione crea regioni dello spazio dei parametri con curvatura di Fisher molto eterogenea: le teste di attenzione focalizzate su poche posizioni (con p concentrata) hanno curvatura locale alta, mentre le teste con attenzione diffusa (p uniforme) hanno curvatura bassa. Questa eterogeneità di curvatura è una delle ragioni per cui l'addestramento dei trasformatori è sensibile alla dimensione del passo e richiede scheduler del learning rate elaborati come il warmup coseno..

Il legame tra la geometria dell'informazione e l'addestramento dei trasformatori è un fronte di ricerca attivo: metodi come Sophia (Liu et al., 2023) e CASPR approssimano la curvatura di Fisher-Gauss-Newton per i transformer, con guadagni in efficienza computazionale rispetto ad Adam documentati empiricamente. Il programma della geometria dell'informazione per i transformer è il tentativo di portare la

Silvano Salvador

giustificazione teorica della geometria riemanniana a un'architettura che domina il campo e il cui addestramento è ancora in larga misura empirico.

A.VI.11 — Sintesi geometrica: un diagramma unificante

La struttura geometrica che attraversa tutte le appendici di questo saggio può essere sintetizzata nel seguente schema di relazioni:

Lo spazio delle politiche π_θ è una varietà statistica $\mathscr{S}$ con metrica di Fisher $G(\theta)$. Il gradiente euclideo $\nabla_\theta J$ indica la direzione di massima crescita di J nella metrica piatta, dipendente dalla parametrizzazione. Il gradiente naturale $G(\theta)^{-1} \nabla_\theta J$ indica la direzione di massima crescita nella metrica di Fisher, invariante per riparametrizzazione.

Le dinamiche di replica sono il gradiente naturale dell'idoneità media sullo spazio delle popolazioni con metrica di Fisher categorica. TRPO è la discesa del gradiente naturale con vincolo geodetico di raggio $\sqrt{(2\delta)}$. PPO approssima TRPO rimpiazzando il vincolo geodetico con il clipping del rapporto di importanza. Il Soft Actor-Critic ottimizza l'energia libera variazionale — la differenza tra ritorno atteso e divergenza KL dalla politica prior — che è equivalente alla massimizzazione dell'ELBO nell'inferenza variazionale.

L'equazione di Hamilton-Jacobi-Bellman è l'equazione delle caratteristiche della funzione valore ottimale nello spazio degli stati fisici; il gradiente naturale è l'analogo sul corrispondente spazio delle politiche. I world models apprendono una varietà latente dello spazio degli stati fisici e trasferiscono l'ottimizzazione della politica dalla varietà fisica a

Silvano Salvador

quella latente, con struttura HJB definita dalla geometria indotta dall'encoder variazionale.

L'allineamento è un problema di distanza nella metrica dello spazio dei valori — una metrica che, per il teorema di Arrow, non può essere ridotta a una forma scalare senza perdere struttura informativa essenziale. L'inner alignment è un problema di fedeltà del trasporto parallelo: il mesa-objective del modello addestrato deve essere il trasporto parallelo del base objective dall'ambiente di addestramento all'ambiente di deployment lungo la geodetica della distribuzione degli stati. Questi due modi di vedere — il primo come problema di Arrow sullo spazio dei valori, il secondo come problema di trasporto parallelo sulla varietà delle politiche — convergono nel medesimo punto: la struttura geometrica è il linguaggio naturale sia dell'ottimizzazione efficiente sia dell'allineamento robusto.

Silvano Salvador